THE UNIVERSITY OF ALBERTA

A LINEAR BOUNDARY COLLOCATION METHOD FOR SOLVING THE DIRICHLET PROBLEM FOR LAPLACE'S EQUATION

by

Clifford G. Morgan ©

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

JULY, 1968

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled A LINEAR BOUNDARY COLLOCATION METHOD FOR SOLVING THE DIRICHLET PROBLEM FOR LAPLACE'S EQUATION submitted by Clifford G. Morgan in partial fulfilment of the requirements for the degree of Master of Science.

ABSTRACT

This thesis presents a computationally-oriented linear boundary collocation technique for obtaining a close approximation to the solution of the Dirichlet problem for Laplace's equation. Necessary concepts and algorithms from the theory of linear approximation are reviewed, and numerical results are given.

ACKNOWLEDGEMENTS

I express my appreciation to my advisor, Dr. S. Charmonman, for his supervision of this research; to Dr. G. Syms, a member of the thesis examining committee, for his reading of the manuscript; to Dr. K.C. Cheng of the Department of Mechanical Engineering, another member of the thesis examining committee, who introduced me to the subject of linear boundary collocation methods when I worked for him as a summer assistant in 1966; and to Dr. A. Sharma of the Department of Mathematics for his interest and remarks.

I also wish to thank the National Research Council of Canada and the Department of Computing Science for financial assistance for carrying out this research.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

This thesis reviews the exchange algorithm for computing a best linear approximation on a discrete point set and shows how it may be used by an algorithm of Remes to compute a best linear approximation on a continmum. These two algorithms are employed by a linear boundary collocation method to compute a close approximation to the solution of the Dirichlet problem for Laplace's equation.

1.1 The Dirichlet Problem for Laplace's Equation

Consider the following classical problem from the theory of partial differential equations: Let C be a simple closed curve, R the region interior to C, and f a continuous function on C. Find a function $u = u(x,y)$ having continuous second order derivatives in R satisfying

$$(1.1) \qquad u_{xx} + u_{yy} = 0 \quad \text{in} \quad R ,$$

and

$$(1.2) \qquad u = f \quad \text{on} \quad C .$$

The subscripts in (1.1) denote partial differentiation.

Equation (1.1) is known as Laplace's equation. Functions having continuous second order derivatives in R and satisfying (1.1) are called harmonic functions in R. Equation (1.2) is the Dirichlet boundary condition for Laplace's equation. Determining the solution u is the Dirichlet problem for Laplace's equation. Ahlfors ([1], pp. 175-199) has proven that u exists and is unique.

The Dirichlet problem arises in a wide variety of physical contexts, including heat transfer, hydrodynamics, electricity, and aerodynamics. Sample problems may be found in Forsythe and Wasow [16] and Fox [17].

Except for some simple boundaries such as the rectangle and the circle, where analytical solutions can be obtained in closed form, most boundaries require that some numerical method be used to obtain the solution. Most of the numerical work devoted to the Dirichlet problem has been concerned with replacing the derivatives in (1.1) by appropriate finite differences and then determining the solution at a finite set of points in the region R. However, this technique is difficult to apply to regions whose boundaries are irregular, and the resulting table of solution values is not convenient for use in subsequent numerical work. Attempts have therefore been made to derive methods which will not only yield the solution in closed form, but also be easy to apply to irregularly shaped boundaries.

1.2 Collocation Methods

Attempts to solve boundary-value problems in partial differential equations are often made using collocation methods. One assumes as an approximation to the solution $u(x,y)$ an expression of the form $g(x,y;c)$ which depends on a vector of parameters $c = (c_1,\ldots,c_n)$ such that for arbitrary values of c

a) the differential equation is already satisfied exactly, or

b) the boundary conditions are already satisfied exactly, or

c) g satisfies neither the differential equation nor the boundary conditions.

One then tries to determine c so that g satisfies

in method (a), the boundary conditions, or

in method (b), the differential equation, or

in method (c), the boundary conditions and the differential equation

as accurately as possible in some sense yet to be defined.

Definition 1.1. *A collocation method of type (a) described previously is called a boundary collocation method.*

For partial differential equations, boundary collocation methods are usually preferred for two reasons. First, it is generally easier to find functions which satisfy a particular differential equation than it is to find functions which

satisfy the boundary conditions for a particular boundary. Second, boundary collocation methods are usually chosen since their use, in so far as integration is concerned, requires the evaluation of integrals around the boundary C rather than throughout the region R. These observations were also verified in practice by Shuleshko [29], who made a comparative analysis of the three collocation techniques applied to the torsion problem and concluded that not only were boundary collocation methods easier to apply, but also yielded better results.

Linear partial differential equations have the property that $\alpha u + \beta v$ is a solution of the partial differential equation if u and v satisfy the equation and α and β are any two constants. Hence boundary collocation methods for linear partial differential equations can have a special form.

Definition 1.2. *A boundary collocation method for a linear partial differential equation which uses an approximating function of the form* $g = \sum_{j=1}^{n} c_j g_j$ *, where each function* g_j *,* $j=1,\ldots,n$ *, satisfies that partial differential equation, is called a linear boundary collocation method.*

1.3 Linear Boundary Collocation Methods for the Dirichlet Problem

Because Laplace's equation is linear, and since the sum of two continuous functions is continuous, the following is a linear boundary collocation method for the Dirichlet problem: Let $g_1,\ldots,g_n$ be harmonic functions in the region R. Then for an arbitrary vector of coefficients $c = (c_1,\ldots,c_n)$, $\sum_{j=1}^{n} c_j g_j$ is also a harmonic function in R. All linear boundary collocation methods are distinguished by the manner in which c is determined.

1.3.1 Interpolation Techniques In this method, n distinct points $s_1,\ldots,s_n$ are selected on the boundary C, and the parameters are chosen so that $\sum_{j=1}^{n} c_j g_j$ interpolates to f at these points. That is, $(c_1,\ldots,c_n)$ is the solution of the system of linear equations,

$$(1.3) \qquad \sum_{j=1}^{n} c_j g_j(s_i) = f(s_i) \; , \quad i=1,\ldots,n \; .$$

The essential idea of this technique was first suggested by Walsh [34] in 1929, when he discussed the problem of approximating harmonic functions by harmonic polynomials. Since then the method has been applied with varying success to a number of engineering problems, including ones in heat transfer, elasticity, and the bending and buckling

of plates [4,5,6,7,8,19,23,26,28,30,33]. The technique was first dubbed the "point-matching method" by Conway [6] in 1961.

Not too much is known about this method theoretically. Its use seems to indicate that its success depends upon the geometry of the problem, but this dependence has not been explicitly determined. Another problem arises from the fact that the coefficient matrix in (1.3) may be singular for some sets of points on C . Further, it is not known under what conditions the method converges. Perhaps the most critical problem is that a general set of properties of the "best" set of matching points has not been found.

Theoretical material concerning this technique is discussed in a series of papers by Curtiss [9-11] and Sobczyk [31]. On the basis of simple cases of circles and ellipses that have been investigated successfully [35,36], and also arguing by analogy from the case of complex analytic interpolation, Curtiss [9] concludes that for sufficiently regular boundaries there probably exist sequences of points for which this interpolation process converges for a wide class of boundary data. However, his suggestion that these points are probably obtainable as the exterior conformal image of equally spaced points on the unit circle (such as the roots of unity) is unsuitable for numerical work because the conformal map is not always

available. Sobczyk [31] has even shown that the resulting image points on C may result in a singular coefficient matrix in (1.3).

1.3.2 Least-Squares Techniques In a least-squares approach, c is chosen to minimize

$$(1.4) \qquad E_1 = \int_C [f(s) - \sum_{j=1}^{n} c_j g_j(s)]^2 \, ds \ .$$

Usually the integration in (1.4) cannot be carried out analytically. Therefore, m points are chosen on the boundary C and m weights w_i corresponding to a quadrature rule, and then the vector c is chosen to minimize

$$(1.5) \qquad E_2 = \sum_{i=1}^{m} w_i [f(s_i) - \sum_{j=1}^{n} c_j g_j(s_i)]^2 \ .$$

In contrast to the interpolation technique, the discrete process in (1.5) should be convergent as n becomes large and $m >> n$. Details of this method and numerical results can be found in papers by Lo [20], Davis [12], Davis and Rabinowitz [14], O'Jalvo and Linzer [24], Merriman [21], and Sparrow [32].

1.4 Purpose of Study

In Chapter II, we shall review some basic theory relevant to the Dirichlet problem and thus attempt to show that choosing c to minimize

$$(1.6) \qquad M = \max_{s \varepsilon C} \left| f(s) - \sum_{j=1}^{n} c_j g_j(s) \right|$$

is a more natural linear boundary collocation method than either the point-matching method or the least-squares method. In Chapter III, IV and V we shall show that such a vector c always exists and we shall present a computationally oriented algorithm for finding it. Results of numerical computations will be given in Chapter VI, and conclusions and suggestions for further research in Chapter VII.

A discrete version of this approach has been made by Quon and Leung [27] using linear programming techniques.

CHAPTER II

BASIC THEORY

2.1 Some Concepts in Linear Approximation Theory

The problem outlined in the previous chapter (see Section 1.4) is a typical one in approximation theory in that it involves the selection from a given set of functions one that is in some sense "close" to a prescribed function (usually) not in the set. Any discussion of approximation requires that we have a way of measuring the "closeness" of two functions. It is reasonable to demand that this measure should have some of the properties of geometrical distance, or more precisely, that it should be a metric.

2.1.1 Metric Spaces

Definition 2.1. *Let S be a set of elements $x, y, z, \ldots,$ and let d be a real-valued function defined on pairs of elements of S. Then d is called a metric if it possesses the following properties:*

$$(2.1)\qquad \begin{aligned} &i) \quad d(x,x) = 0, \\ &ii) \quad d(x,y) > 0 \quad \text{if} \quad x \neq y, \\ &iii) \quad d(x,y) = d(y,x), \\ &iv) \quad d(x,y) \leq d(x,z) + d(z,y). \end{aligned}$$

The ordered pair (S,d) is then termed a metric space.

If, for example, S is the set of real numbers, the usual metric is $d(x,y) = |x-y|$. Properties (i) to (iv) are then simply familiar properties of the absolute value function.

The concept of a metric space is usually too general to yield many practical results and therefore the setting for many problems in the theory of approximation is a metric space of a particular kind called a normed linear space. We must first, however, define a linear space.

2.1.2 Linear Spaces

Definition 2.2. *Let S be a set of elements $x,y,z,\ldots$ for which two types of operations are possible. S is termed a linear space if any two elements x,y in S determine a unique element $x+y$ in S as their sum, if each element x in S and each scalar α determines a unique element αx in S as their scalar product, and if summation and scalar multiplication satisfy the following properties:*

i) $x+y = y+x$,

ii) $x + (y+z) = (x+y) + z$,

iii) There exists a unique element Θ in S such that $x+\Theta = x$ for all x in S (Θ is called the zero element of the space),

(2.2) iv) *For each x in S, there exists a unique inverse $(-x)$ in S such that $x + (-x) = \Theta$,*

v) $\alpha(\beta x) = (\alpha\beta)x$ *for all scalars* α,β,

vi) $\alpha(x+y) = \alpha x + \alpha y$,

vii) $(\alpha+\beta)x = \alpha x + \beta x$,

viii) $1x = x$.

We now list some linear spaces which are important to this study.

<u>Example 2.1</u>. *The real, n-dimensional, Euclidian space, denoted by* R_n .

R_n consists of vectors $x,y,z,\ldots$ that are n-tuples of real numbers: $x = (x_1,\ldots,x_n)$. Summation is defined by $x+y = (x_1+y_1,\ldots,x_n+y_n)$, and scalar multiplication by $\alpha x = (\alpha x_1,\ldots,\alpha x_n)$. The zero element is $\Theta = (0,\ldots,0)$, and the inverse of x is $(-x) = (-x_1,\ldots,-x_n)$.

<u>Example 2.2</u>. *The space of all functions continuous in the interval* $[a,b]$, *denoted by* $C[a,b]$.

Let $f,g,\ldots$ be members of $C[a,b]$ and let x be any point in $[a,b]$. We define the sum of f and g by $(f+g)(x) = f(x) + g(x)$, and the scalar product of α and f by $(\alpha f)(x) = \alpha f(x)$. The zero element is that function which vanishes identically in $[a,b]$ and the inverse of f is given by $(-f)(x) = -f(x)$.

Example 2.3. *The space of all functions continuous on a simple closed curve* C *, denoted by* $W(C)$.

Example 2.4. *Let* C *be a simple closed curve and* R *the region interior to* C . *The space of all complex functions analytic in* R *and continuous on* C *, denoted by* $A(R,C)$.

Before leaving the concept of a linear space, we shall present a definition which will be used in subsequent work.

Definition 2.3. *Let* S *be a linear space. A subset* M *of* S *is called convex if* x,y *are any two distinct elements of* M *implies that* $\alpha x + (1-\alpha)y$ *is also a member of* M *for any real* α , $0 \leq \alpha \leq 1$.

If we picture M geometrically, then $\alpha x + (1-\alpha)y$ lies on the line segment joining x and y . Thus M is convex if it contains all the points on the line segment joining any two of its points. Circles, ellipses, rectangles, and straight lines are examples of convex sets in R_n . The figure "8" is not convex in R_n .

2.1.3 Normed Linear Spaces

Definition 2.4. *A normed linear space is a linear space* S *for which there is defined a real-valued function on elements* $x,y,z,\ldots$ *in* S *called a norm, denoted by* $\|x\|$ *, and satisfying the following laws:*

$$(2.3)\qquad \begin{aligned} &i)\quad \|\Theta\| = 0,\\ &ii)\quad \|x\| > 0 \quad if \quad x \neq \Theta,\\ &iii)\quad \|\alpha x\| = |\alpha|\ \|x\| \quad for\ any\ scalar \quad \alpha,\\ &iv)\quad \|x+y\| \le \|x\| + \|y\| \ . \end{aligned}$$

An easily verified property of norms is given by

Theorem 2.1. $\big|\|x\| - \|y\|\big| \le \|x-y\|$.

There are usually an infinity of norms that may be introduced into linear spaces in order to make them normed linear spaces. However, we discuss only those relevant to this study.

Theorem 2.2. *The linear space* R_n *with norm defined by*

$$(2.4)\qquad \|x\| = \max_{1 \le i \le n} |x_i|$$

is a normed linear space.

Proof: We must verify that the function given in (2.4) satisfies the properties listed in (2.3). The first three properties are easily verified as simple consequences of properties of the absolute value function. To prove (iv) we write

$$\begin{aligned}\|x+y\| &= \|(x_1+y_1,\ldots,x_n+y_n)\| \\ &= \max_{1\le i\le n} |x_i+y_i| \\ &= |x_k+y_k| \text{ , for some } k \text{ , } 1\le k\le n \\ &\le |x_k| + |y_k| \\ &\le \max_{1\le i\le n} |x_i| + \max_{1\le i\le n} |y_i| \\ &= \|x\| + \|y\| \text{ .}\end{aligned}$$

Theorem 2.3. *The linear space $C[a,b]$ with norm defined by*

$$\|f\| = \max_{a\le x\le b} |f(x)| \tag{2.5}$$

is a normed linear space.

Theorem 2.4. *The linear space $W(C)$ with norm defined by*

$$\|f\| = \max_{(x,y)\varepsilon C} |f(x,y)| \tag{2.6}$$

is a normed linear space.

Theorem 2.5. *The linear space $A(R,C)$ with norm defined by*

(2.7) $$\|f\| = \max_{z\varepsilon C} |f(z)|$$

is a normed linear space.

Proof: The maximum modulus need only be taken over C, since a complex function analytic in R and continuous on C assumes its maximum modulus on C ([25], pp. 290-291).

Theorem 2.6. *In a normed linear space*

(2.8) $$d(f,g) = \|f-g\|$$

defines a metric.

Proof: We must show that (2.3) and (2.8) imply (2.1).

i) $d(f,f) = \|f-f\| = \|0 \cdot f\| = |0| \, \|f\| = 0$.

ii) $d(f,g) = \|f-g\| > 0$ since $f \neq g$ implies $f-g \neq \Theta$.

iii) $d(f,g) = \|f-g\| = |-1| \, \|f-g\| = \|g-f\| = d(g,f)$

iv) $d(f,g) = \|f-g\| = \|f-h+h-g\| \leq \|f-h\| + \|h-g\|$

$= d(f,h) + d(h,g)$.

2.1.4 Best Approximation From the remarks made at the beginning of this Section and from Theorem 2.6, we see that a meaningful way of defining the closeness of two functions f and g in a normed linear space is to define it as the norm of their difference.

Definition 2.5. *Let* $f, g_1, \ldots, g_n$ *be elements of a normed linear space* S . *A best approximation to* f *by linear combinations of* $g_1, \ldots, g_n$ *is an element* $\sum_{j=1}^{n} c_j g_j$ *for which*

$$(2.9) \qquad \|f - \sum_{j=1}^{n} c_j g_j\| \leq \|f - \sum_{j=1}^{n} a_j g_j\|$$

for every choice of constants $a_1, \ldots, a_n$.

Then the fundamental assertion of linear approximation theory is

Theorem 2.7. *The problem of finding* $\min_{a_j} \|f - \sum_{j=1}^{n} a_j g_j\|$ *has a solution.*

A proof of this important theorem can be found in ([13], pp. 137-139).

Definition 2.6. *For a given* $f, g_1, \ldots, g_n$ *in a normed linear space,* $E_n(f)$ *is defined by*

$$(2.10) \qquad E_n(f) = \min_{a_j} \|f - \sum_{j=1}^{n} a_j g_j\| .$$

Theorem 2.8.

$$E_1(f) \geq E_2(f) \geq \ldots \tag{2.11}$$

Proof: (2.11) is true since linear combinations of $g_1, \ldots, g_n$ are also linear combinations of $g_1, \ldots, g_n, g_{n+1}$.

We have observed there is always one best approximation to f by linear combinations $\sum_{j=1}^{n} a_j g_j$. However, there may be more than one. If M is the set of best approximations, then we have

Theorem 2.9. *M is a convex set.*

A proof of the theorem can be found in ([13], pp. 140-141).

In our discussion of best approximation, the particular norm used was not specified. However, since the norms given in Theorems 2.2-2.5 are formulated in terms of a certain maximum, the corresponding best approximations solve the problem of minimizing the maximum error, and therefore we will subsequently refer to these best approximations as minimax approximations.

2.2 Some Properties of Harmonic Functions

Laplace's equation bears a close relationship to the theory of analytic functions. In fact, we have

Theorem 2.10. *The real part of any analytic function is a harmonic function; and conversely, every*

harmonic function is the real part of some analytic function.

Proof: Set $z = x + iy$ and let $f(z) = u(x,y) + iv(x,y)$ be an analytic function of z. Then u and v are related by the Cauchy-Riemann equations

$$u_x = v_y \ , \ u_y = -v_x \ , \tag{2.12}$$

and are said to be conjugate ([25], p. 83). Furthermore, u and v possess continuous partial derivatives of all orders ([25], pp. 185-186). If the first of these equations is differentiated with respect to x, the second with respect to y, and the two resulting equations added, then it can be seen that $u(x,y) = \text{Re } f(z)$ is a solution of Laplace's equation. On the other hand, let u be a harmonic function. Then equations (2.12) can be solved for v, and since the partial derivatives u_x, u_y, v_x, v_y are continuous, $f(z) = u(x,y) + iv(x,y)$ is an analytic function of z ([25], p. 85).

A parallel argument yields a similar theorem for the imaginary parts of analytic functions.

Another very important property of harmonic functions is given in

Theorem 2.11. *Let C be a simple closed curve and R the region interior to C. Let u be a harmonic*

function in R *and continuous on* C *. Then there exist two points* z_1 *and* z_2 *on* C *such that* $u(z_1) \le u(z) \le u(z_2)$ *for all* z *in* R *.*

A proof of this theorem can be found in ([25], pp. 349-350). A crucial result of this theorem is

Theorem 2.12. *Consider Dirichlet's problem for Laplace's equation (see Section 1.1). If* $g(x,y)$ *is a harmonic function such that*

$$\max_{(x,y)\varepsilon C} |f(x,y) - g(x,y)| \le M \tag{2.13}$$

for some constant M *, then*

$$\max_{(x,y)\varepsilon R} |u(x,y) - g(x,y)| \le M \ . \tag{2.14}$$

Proof: If u is the solution of the original problem, then u-g is the solution of the same problem with boundary function f-g rather than f . Note that f-g is continuous and u-g is harmonic. However, by Theorem 2.11, a harmonic function assumesits maximum value on the boundary C . That is,

$$\max_{(x,y)\varepsilon R} |u(x,y) - g(x,y)| \le \max_{(x,y)\varepsilon C} |f(x,y) - g(x,y)| \le M \ .$$

2.3 A Complete Set of Harmonic Functions

So far we have not specifically given a set of harmonic functions $g_1,\ldots,g_n$ with which to approximate u in R and f on C. We see, from Theorem 2.10, that we need only use the real and/or imaginary parts of some analytic functions. Now z^n, $n=0,1,\ldots$ are obviously analytic functions and hence

$$H = \{1, \operatorname{Re} z^n, \operatorname{Im} z^n, n=1,2,\ldots\} \tag{2.15}$$

is an infinite set of harmonic functions. They are commonly called harmonic polynomials.

Once a particular set of harmonic functions have been chosen, a problem arises as to the "completeness" of this set. Completeness refers to the possibility of arbitrarily close approximation of the boundary data by means of the boundary values of linear combinations of the selected set of harmonic functions. A relevant result, due to J.L. Walsh, is that the powers, $1,z,z^2,\ldots$ are complete in $A(R,C)$ ([13], pp. 275-278). An important consequence of Walsh's assertion is

Theorem 2.13. *The harmonic polynomials are complete in $W(C)$.*

Proof: Let u be the solution of the Dirichlet problem. Then u is the real part of some analytic function $F(z)$, by Theorem 2.10. By the completeness of the powers $1, z, z^2, \ldots$ in $A(R,C)$, there exists a complex polynomial $p_n(z)$ such that $\max\limits_{z \varepsilon C} |F(z) - p_n(z)| \leq \delta$ for any $\delta > 0$. We now write

$$\begin{aligned} \delta &\geq \max_{z \varepsilon C} |F(z) - p_n(z)| \\ &\geq |F(z) - p_n(z)| \quad \text{for any } z \text{ on } C \\ &\geq |\mathrm{Re}\ (F(z) - p_n(z))| \\ &\geq |f(x,y) - \mathrm{Re}\ p_n(z)| \end{aligned}$$

Since the above inequality is true for any z on C, we have that

$$\delta \geq \max_{(x,y) \varepsilon C} |f(x,y) - \mathrm{Re}\ p_n(z)| .$$

Letting δ tend to zero proves the theorem.

CHAPTER III

MINIMAX APPROXIMATIONS ON A DISCRETE POINT SET

3.1 Introduction

Given the Dirichlet problem outlined in Chapter I (see Section 1.1), we set ourselves the task of selecting a vector of parameters $c = (c_1,\ldots,c_n)$ to minimize the expression

$$(3.1) \qquad M_1 = \max_{(x,y)\varepsilon C} \left|f(x,y) - \sum_{j=1}^{n} a_j g_j(x,y)\right| .$$

Our discussions in Chapter II ensured us that such a vector c existed. Furthermore, if g_j , $j=1,\ldots,n$ were chosen to be the harmonic polynomials, then for sufficiently large n , $\sum_{j=1}^{n} c_j g_j$ could be accepted as a reasonably good approximate solution to the Dirichlet problem.

As yet, however, we have not tackled the problem of <u>how</u> to determine the vector c . As a first step, we might select m points (x_i,y_i) , $i=1,\ldots,m$ $(m>n)$ from the boundary C and minimize the expression

$$(3.2) \qquad M_2 = \max_{1\le i\le m} \left|f(x_i,y_i) - \sum_{j=1}^{n} a_j g_j(x_i,y_i)\right| ,$$

accepting the minimax solution to (3.2) as a close approximation to the minimax solution of (3.1). This is equivalent to solving the following linear system of equations in the minimax sense:

$$(3.3) \qquad \sum_{j=1}^{n} a_j g_j(x_i,y_i) = f(x_i,y_i) \ , \quad i=1,\ldots,m \ .$$

Obtaining a minimax solution to an overdetermined, inconsistent system of linear equations is an important problem in its own right. The discussion we now give parallels that given by Handscomb ([18], pp. 73-82). A more advanced and complete approach has been given by Cheney ([3], pp. 28-56).

3.2 Minimax Solution of an Overdetermined, Inconsistent System of Linear Equations

Consider the system of linear equations

$$(3.4) \qquad \sum_{j=1}^{n} a_{ij}x_j = b_i \ , \quad i=1,\ldots,m \ .$$

We assume that the numbers a_{ij} and b_i are given and that the unknowns x_j are to be determined. (3.4) is frequently written in matrix notation as $Ax = b$. The system may have exactly one solution, infinitely many

solutions, or no solution, depending on the data. If the rank of the coefficient matrix A is equal to the rank of the augmented matrix A_b, then (3.4) possesses a solution and is said to be consistent. If $m>n$ and the rank of A does not equal the rank of A_b, then (3.4) is said to be overdetermined and inconsistent. However, even though (3.4) may possess no solution, we may still try to find a vector $x = (x_1,\ldots,x_n)$ that minimizes the expression

$$\rho = \max_{1\le i\le m} \left| b_i - \sum_{j=1}^{n} a_{ij}x_j \right| . \tag{3.5}$$

Such an x is called a minimax solution of (3.4).

3.2.1 Some Properties of Minimax Solutions

Theorem 3.1. *Let values a_{ij}, b_i be given for $j=1,\ldots,n$ and $i=1,\ldots,m>n$. The problem of finding*

$$\min_{x\in R_n} \max_{1\le i\le m} \left| b_i - \sum_{j=1}^{n} a_{ij}x_j \right|$$

has a solution.

Proof: For each $j=1,\ldots,n$, let A_j denote the column vector $(a_{1j},\ldots,a_{mj})^T$. Then (3.4) can be rewritten

in the form

$$\sum_{j=1}^{n} x_j A_j = b \ . \tag{3.6}$$

Now $\sum_{j=1}^{n} y_j A_j$ and b are elements of R_m for any $y = (y_1,\ldots,y_n)$. Hence, by Theorems 2.2 and 2.7, there exists a vector $x \varepsilon R_n$ such that

$$\left\| b - \sum_{j=1}^{n} x_j A_j \right\| \leq \left\| b - \sum_{j=1}^{n} y_j A_j \right\| \tag{3.7}$$

for any $y \varepsilon R_n$.

Theorem 3.2. *The set of minimax xolutions to (3.4) is convex.*

Proof: Theorem 3.2 is just a rewording of Theorem 2.9.

Before proceeding further, we make the following important definition.

Definition 3.1. *A set of vectors in R_n is said to satisfy Haar's condition if every subset of n of them is linearly independent.*

We now make the rather strong assumption that the set of row vectors making up the matrix A satisfy Haar's condition. Therefore, any n of the m equations in (3.4)

possess a nonsingular coefficient matrix. Haar's condition will be discussed in detail in Chapter IV.

Theorem 3.3. *Let $x = (x_1, \ldots, x_n)$ be a minimax solution to (3.4) and let $\rho = \max_{1 \le i \le m} |b_i - \sum_{j=1}^{n} a_{ij}x_j|$. Then $I = \{i : |b_i - \sum_{j=1}^{n} a_{ij}x_j| = \rho\}$ contains at least $(n+1)$ elements.*

Proof: Suppose, on the contrary, that I contains r elements, where $r < n+1$. Reordering the equations if necessary, we can assume that the maximum absolute error occurs in the first r equations. Thus we have

$$b_1 - \sum_{j=1}^{n} a_{1j}x_j = s_1\rho$$

$$\cdots\cdots\cdots\cdots$$

$$b_r - \sum_{j=1}^{n} a_{rj}x_j = s_r\rho$$

(3.8)

$$b_{r+1} - \sum_{j=1}^{n} a_{r+1,j}x_j = s_{r+1}e_{r+1}$$

$$\cdots\cdots\cdots\cdots$$

$$b_m - \sum_{j=1}^{n} a_{mj}x_j = s_me_m$$

where

$$(3.9) \qquad 0 \le e_i < \rho \ , \quad i=r+1,\dots,m \ ,$$

and

$$(3.10) \qquad s_i = \operatorname{sgn}(b_i - \sum_{j=1}^{n} a_{ij}x_j) \ , \quad i=1,\dots,m \ .$$

We now show that the rank of the matrix $A_r = (a_{ij})$, $j=1,\dots,n$; $i=1,\dots,r$; is less than r. If, on the other hand, the rank of A_r equals r, then we can find a vector $u = (u_1,\dots,u_n)$ such that

$$(3.11) \qquad \sum_{j=1}^{n} a_{ij}u_j = s_i\rho \ , \quad i=1,\dots,r \ .$$

Now define a new vector $y = (y_1,\dots,y_n)$ by

$$(3.12) \qquad y_j = x_j + \lambda u_j \ , \quad j=1,\dots,n \ ,$$

where

$$(3.13) \qquad 0 < \lambda < \min(\frac{\rho-M_1}{M_2} \ , \ 1) \ .$$

M_1 and M_2 in (3.13) are given by

(3.14)
$$M_1 = \max_{r+1 \le i \le m} e_i ,$$

and

(3.15)
$$M_2 = \max_{r+1 \le i \le m} \left| \sum_{j=1}^{n} a_{ij} u_j \right| .$$

Then for $i=1,\dots,m$ we have

$$b_i - \sum_{j=1}^{n} a_{ij} y_j = b_i - \sum_{j=1}^{n} a_{ij}(x_j + \lambda u_j)$$

$$= b_i - \sum_{j=1}^{n} a_{ij} x_j - \lambda \sum_{j=1}^{n} a_{ij} u_j$$

(3.16)
$$= \begin{cases} s_i \rho - \lambda\, s_i \rho , & i=1,\dots,r . \\ s_i \rho_i - \lambda \sum_{j=1}^{n} a_{ij} u_j , & i=r+1,\dots,m . \end{cases}$$

Taking the absolute value of (3.16) we get

(3.17)
$$|b_i - \sum_{j=1}^{n} a_{ij}y_j| = \begin{cases} (1-\lambda)\rho \ , & i=1,\ldots,r \ . \\ |s_ie_i - \lambda \sum_{j=1}^{n} a_{ij}u_j| \ , & i=r+1,\ldots,m \ . \end{cases}$$

Certainly, $(1-\lambda)\rho < \rho$ and for $i=r+1,\ldots,m$

(3.18)
$$\begin{aligned} |s_ie_i - \lambda \sum_{j=1}^{n} a_{ij}u_j| &\le e_i + \lambda \,|\sum_{j=1}^{n} a_{ij}u_j| \\ &< e_i + \frac{\rho-M_1}{M_2} \,|\sum_{j=1}^{n} a_{ij}u_j| \\ &\le e_i + \rho - M_1 \ , \text{ by (3.15)} \\ &\le \rho \ , \text{ by (3.14).} \end{aligned}$$

This means that y is a better minimax approximation than x, which is a contradiction. Hence, the rank of A_r is less than r. Since $r \le n$, this contradicts our basic assumption that the rows of A satisfy Haar's condition. Hence, I contains at least $(n+1)$ elements.

Theorem 3.4. *The minimax solution to (3.4) is unique.*

Proof: Suppose that $x, y \in R_n$ are two distinct minimax solutions and let

$$(3.19)\quad \rho = \max_{1\le i\le m} |b_i - \sum_{j=1}^{n} a_{ij}x_j| = \max_{1\le i\le m} |b_i - \sum_{j=1}^{n} a_{ij}y_j| .$$

Now define

$$(3.20)\qquad v = \tfrac{1}{2}(x+y) .$$

Then v is also a minimax solution by Theorem 3.2. For those i for which

$$(3.21)\qquad b_i - \sum_{j=1}^{n} a_{ij}v_j = s_i\rho ,$$

where s_i is still given by (3.10), we must have

$$(3.22)\quad (b_i - \sum_{j=1}^{n} a_{ij}x_j) + (b_i - \sum_{j=1}^{n} a_{ij}y_j) = 2s_i\rho .$$

Therefore, we obtain

$$(3.23) \qquad 2\rho \le |b_i - \sum_{j=1}^{n} a_{ij}x_j| + |b_i - \sum_{j=1}^{n} a_{ij}y_j| .$$

Neither term on the right of the above inequality can be greater than ρ . Therefore,

$$(3.24) \qquad 2\rho = |b_i - \sum_{j=1}^{n} a_{ij}x_j| + |b_i - \sum_{j=1}^{n} a_{ij}y_j| .$$

By the same reasoning, we obtain

$$(3.25) \qquad \rho = |b_i - \sum_{j=1}^{n} a_{ij}x_j| = |b_i - \sum_{j=1}^{n} a_{ij}y_j| .$$

Hence,

$$(3.26) \qquad b_i - \sum_{j=1}^{n} a_{ij}x_j = b_i - \sum_{j=1}^{n} a_{ij}y_j ,$$

for if the two terms were of opposite sign, this would contradict (3.22). Therefore,

$$(3.27) \qquad \sum_{j=1}^{n} a_{ij}(x_j - y_j) = 0 .$$

By Theorem 3.3, there are (n+1) such i for which (3.27) holds. Choose any n of them. The above equation is then a system of n equations in n unknowns. Our basic assumption ensures that the coefficient matrix of this system is nonsingular. Hence, (3.27) has only the zero solution. That is,

$$x_j - y_j = 0 \ , \quad j=1,\ldots,n \ , \tag{3.28}$$

and therefore x=y as was to be shown.

Now that we have some insight into the existence, characterization, and uniqueness of the minimax solution to (3.4), we must approach the problem of actually computing it. We first turn our attention to the special case where the number of equations is one greater than the number of unknowns.

3.2.2 <u>The Minimax Solution of (n+1) Equations in n Unknowns</u> From Theorem 3.3, we know the minimax solution yields the same absolute error in each equation. Call this absolute error ρ . Using the notation $s_i = \pm 1$ for the sign of the error in the i^{th} equation, we have that the minimax solution x satisfies

$$b_i - \sum_{j=1}^{n} a_{ij}x_j = s_i\rho \ , \quad i=1,\ldots,n+1 \ . \tag{3.29}$$

For a particular choice of signs, the above system can be rewritten as

$$(3.30)\qquad \sum_{j=1}^{n} a_{ij}x_j + s_i\rho = b_i \ , \quad i=1,\dots,n+1 \ ,$$

which is simply a system of $(n+1)$ equations in the $(n+1)$ unknowns $x_1,\dots,x_n,\rho$. Let us define

$$(3.31)\qquad T_i = (-1)^i \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots \\ a_{n+1,1} & \cdots & a_{n+1,n} \end{vmatrix} \ , \quad i=1,\dots,n+1 \ .$$

Note that by Haar's condition, $T_i \neq 0$ for all i . We then have

<u>Theorem 3.5</u>. *The signs of* $s_1,\dots,s_{n+1}$ *are the same as or opposite to the corresponding signs of* $T_1,\dots,T_{n+1}$.

Proof: The coefficient matrix of (3.30) is given by

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} & s_1 \\ \vdots & & \vdots & \vdots \\ a_{n+1,1} & \cdots & a_{n+1,n} & s_{n+1} \end{bmatrix}$$

Solving for ρ by Cramer's rule, expanding determinants down the last column, and utilizing (3.31), we get

$$(3.32) \qquad \rho = \left(\sum_{i=1}^{n+1} T_i b_i\right) \Big/ \sum_{i=1}^{n+1} s_i T_i \ .$$

We are trying to determine s_i so that ρ is minimized. Hence, we must try to make the denonimator in (3.32) as large in absolute value as possible. This will happen if

$$(3.33) \qquad s_i = \operatorname{sgn} T_i \ , \quad i=1,\ldots,n+1 \ ,$$

or

$$(3.34) \qquad s_i = - \operatorname{sgn} T_i \ , \quad i=1,\ldots,n+1 \ ,$$

which was to be shown.

Now that the signs s_i have been independently determined, it remains to be shown that the coefficient matrix of (3.30) is nonsingular.

Theorem 3.6

$$D = \begin{vmatrix} a_{11} & \cdots & a_{1n} & s_1 \\ \vdots & & \vdots & \vdots \\ a_{n+1,1} & \cdots & a_{n+1,n} & s_{n+1} \end{vmatrix} \neq 0 \ . \tag{3.35}$$

Proof: We have already remarked that each $T_i \neq 0$. To prove the assertion, we expand the determinant down the last column, obtaining

$$D = \sum_{i=1}^{n+1} s_i T_i \ . \tag{3.36}$$

But by Theorem 3.5, we have chosen s_i according to (3.33) or (3.34), and in one case, we have a sum of all positive terms, and in the other, a sum of all negative terms. In either case,

$$D \neq 0 \ , \tag{3.37}$$

and the theorem is established.

To summarize, we can obtain the minimax solution of $(n+1)$ equations in n unknowns by first determining the numbers s_i according to (3.33) or (3.34) and then solving the linear system (3.30) for $x_1,\ldots,x_n,\rho$.

We now return to the original problem of determining the minimax solution to (3.4). The results just presented are important because the minimax solution of (3.4) is also the minimax solution to some subsystem of (3.4) consisting of just $(n+1)$ equations (see Theorem 3.3). The main problem now, then, is to determine which subset of $(n+1)$ equations is pertinent. This subset can be determined by the exchange algorithm.

3.2.3 The Exchange Algorithm The basic idea of this algorithm is to calculate the minimax solutions of a succession of subsystems, each consisting of $(n+1)$ equations from (3.4). By Theorem 3.3, the minimax solution of one of these subsystems is the required minimax solution. Since there are only a finite number of possible subsystems, the exchange algorithm will converge in a finite number of steps, providing we can show that it never reconsiders a previous subsystem.

3.2.3.1 Theory of the Exchange Algorithm

Initially, we choose any $(n+1)$ equations from (3.4). Without loss of generality, we may assume that we have

selected the first $(n+1)$. Then form the T_i and s_i according to

$$(3.38) \qquad T_i = (-1)^i \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots \\ a_{n+1,1} & \cdots & a_{n+1,n} \end{vmatrix} , \quad i=1,\ldots,n+1 ,$$

and

$$(3.39) \qquad s_i = \operatorname{sgn} T_i , \quad i=1,\ldots,n+1 .$$

The values of the $(n+1)$ unknowns $x_1,\ldots,x_n,\rho$ are obtained by solving

$$(3.40) \qquad \sum_{j=1}^{n} a_{ij}x_j + s_i\rho = b_i , \quad i=1,\ldots,n+1 .$$

We may at this point assume that $\rho > 0$. If (3.40) gives $\rho < 0$, we can simply replace each s_i by $-s_i$ and each T_i by $-T_i$. We now compute the discrepancies in each

equation from

$$(3.41) \qquad \rho_i = |b_i - \sum_{j=1}^{n} a_{ij}x_j| \ , \quad i=1,\dots,m \ ,$$

and determine $\max_{1\le i\le m} \rho_i$. A test is made to see if $\max_{1\le i\le m} \rho_i > \rho$. If not, then x is our required minimax solution and ρ is the associated minimax error. On the other hand, if the above condition is true, then we may assume that $\max_{1\le i\le m} \rho_i = \rho_{n+2}$, so that

$$(3.42) \quad \sum_{j=1}^{n} a_{n+2,j}x_j + s_{n+2}\rho_{n+2} = b_{n+2} \ , \ \rho_{n+2} > \rho \ ,$$

where, as usual, s_{n+2} is the sign of the error in the (n+2)nd equation.

Theorem 3.7. *If (3.42) holds, then the minimax error* ρ' *of some subsystem consisting of* $(n+1)$ *equations selected from equations* $1,\dots,n+2$ *is greater than* ρ .

The method of proof will be to explicitly give the indices of the equations comprising the relevant subsystem.

Proof: We define a set of row vectors by

$$(3.43) \qquad A_i = (a_{i1},\dots,a_{in}) \ , \quad i=1,\dots,n+2 \ .$$

Now consider the determinants given by

$$(3.44) \qquad D_j = \begin{vmatrix} a_{1j} & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n+1,j} & a_{n+1,1} & \cdots & a_{n+1,n} \end{vmatrix}, \quad j=1,\ldots,n \ .$$

Obviously, $D_j = 0$ for all j since the 1^{st} and $(j+1)^{st}$ columns of the determinant are identical. If we expand D_j down the first column, we obtain

$$(3.45) \qquad 0 = \sum_{i=1}^{n+1} a_{ij}T_i \ , \quad j=1,\ldots,n \ ,$$

which may be rewritten as

$$(3.46) \qquad \sum_{i=1}^{n+1} T_iA_i = 0 \ .$$

Furthermore, since $s_i^2 = 1$, we may rewrite (3.46) as

$$(3.47) \qquad \sum_{i=1}^{n+1} T_is_i(s_iA_i) = 0 \ .$$

If we consider (3.47) as a linear combination of the vectors s_iA_i, then, by (3.39), all the coefficients are positive.

Since no $T_i = 0$, we may solve for any A_k from (3.46), getting

$$A_k = \sum_{\substack{i=1 \\ i \neq k}}^{n+1} -\frac{T_i}{T_k} A_i \, . \tag{3.48}$$

Now by Haar's condition, we know that the vectors $A_1, \ldots, A_{n+1}$ span R_n. Therefore, there exist numbers λ_i such that

$$A_{n+2} = \sum_{i=1}^{n+1} \lambda_i A_i \, . \tag{3.49}$$

Hence,

$$0 = A_{n+2} - \sum_{i=1}^{n+1} \lambda_i A_i$$

$$= A_{n+2} - \lambda_k A_k - \sum_{\substack{i=1 \\ i \neq k}}^{n+1} \lambda_k A_k$$

$$= A_{n+2} - \lambda_k \sum_{\substack{i=1 \\ i \neq k}}^{n+1} -\frac{T_i}{T_k} A_i - \sum_{\substack{i=1 \\ i \neq k}}^{n+1} \lambda_i A_i \tag{3.50}$$

$$= A_{n+2} + \sum_{\substack{i=1 \\ i \neq k}}^{n+1} (\lambda_k \frac{T_i}{T_k} - \lambda_i) A_i$$

$$= s_{n+2} A_{n+2} + \sum_{\substack{i=1 \\ i \neq k}}^{n+1} s_{n+2} s_i (\lambda_k \frac{T_i}{T_k} - \lambda_i) s_i A_i \, .$$

The last line comes from multiplying the previous one by s_{n+2} and again noting that $s_i^2 = 1$. The above equation expresses the fact that we have found numbers such that a linear combination of $s_i A_i$, $i=1,\ldots,n+2;\ i\neq k$ with these numbers as coefficients is equal to zero. Referring to the discussion after (3.47), we must have all of these coefficients positive. That is, we require

$$(3.51) \qquad s_{n+2} s_i (\lambda_k \frac{T_i}{T_k} - \lambda_i) > 0 \ , \quad i=1,\ldots,n+1;\ i\neq k \ .$$

Therefore, we must have

$$(3.52) \qquad s_{n+2} s_i T_i \frac{\lambda_k}{T_k} > s_{n+2} s_i \lambda_i \ , \quad i=1,\ldots,n+1;\ i\neq k \ .$$

From (3.39), we have $s_i T_i > 0$. Hence,

$$(3.53) \qquad s_{n+2} \frac{\lambda_k}{T_k} > s_{n+2} \frac{\lambda_i}{T_i} \ , \quad i=1,\ldots,n+1;\ i\neq k \ ,$$

and therefore k is the index satisfying

$$(3.54) \qquad s_{n+2} \frac{\lambda_k}{T_k} = \max_{1\leq i\leq n+1} s_{n+2} \frac{\lambda_i}{T_i} \ .$$

The index k thus chosen is unique, for otherwise one of the coefficients in (3.50) would vanish, contradicting Haar's condition. It is now asserted that the minimax error ρ' associated with equations $1,\ldots,k-1,k+1,\ldots,n+2$ is greater than the minimax error ρ associated with equations $1,\ldots,n+1$. Let x' be the minimax solution of equations $1,\ldots,k-1,k+1,\ldots,n+2$. Then

$$(3.55) \qquad |b_i - \sum_{j=1}^{n} a_{ij}x_j| < |b_{n+2} - \sum_{j=1}^{n} a_{n+2,j}x_j| ,$$

$$i=1,\ldots,n+1 ,$$

and

$$(3.56) \qquad |b_i - \sum_{j=1}^{n} a_{ij}x'_j| = |b_{n+2} - \sum_{j=1}^{n} a_{n+2,j}x'_j| ,$$

$$i=1,\ldots,n+1;\ i\neq k .$$

Hence, $x \neq x'$, or

$$(3.57) \qquad x - x' \neq 0 .$$

Now we can write

$$s_i \sum_{j=1}^{n} a_{ij}(x_j'-x_j) = s_i(b_i - \sum_{j=1}^{n} a_{ij}x_j)$$

$$- s_i(b_i - \sum_{j=1}^{n} a_{ij}x_j') \tag{3.58}$$

$$= \rho - \rho' \ , \quad i=1,\dots,n+1;\ i\neq k \ .$$

Now by Haar's condition and (3.57), we have

$$\rho - \rho' \neq 0 \ . \tag{3.59}$$

We also have that

$$s_{n+2} \sum_{j=1}^{n} a_{n+2,j}(x_j'-x_j) = s_{n+2}(b_{n+2} - \sum_{j=1}^{n} a_{n+2,j}x_j)$$

$$- s_{n+2}(b_{n+2} - \sum_{j=1}^{n} a_{n+2,j}x_j') \tag{3.60}$$

$$> \rho - \rho' \ .$$

We are trying to show that $\rho - \rho' < 0$. We have already shown that $\rho - \rho' \neq 0$. Suppose, then, that $\rho - \rho' > 0$. Then by (3.58) and (3.60), we can write

$$(3.61) \qquad s_i \sum_{j=1}^{n} a_{ij}(x_j'-x_j) > 0 \ , \quad i=1,\dots,n+2;\ i\neq k \ .$$

From (3.50) and (3.54), we know there exist numbers μ_i such that

$$(3.62) \qquad \sum_{\substack{i=1\\ i\neq k}}^{n+2} \mu_i s_i A_i = 0 \ ,$$

or equivalently,

$$(3.63) \qquad \sum_{\substack{i=1\\ i\neq k}}^{n+2} \mu_i s_i a_{ij} = 0 \ , \quad j=1,\dots,n \ .$$

Since the $\mu_i > 0$, we can rewrite (3.61) as

$$(3.64) \qquad \mu_i s_i \sum_{j=1}^{n} a_{ij}(x_j'-x_j) > 0 \ , \quad i=1,\dots,n+2;\ i\neq k \ .$$

For any numbers y_j , (3.63) can be altered to read

$$(3.65) \qquad y_j \sum_{\substack{i=1 \\ i\neq k}}^{n+2} \mu_i s_i A_i = 0 \ , \quad j=1,\ldots,n \ .$$

Summing the last set of equations over j , we obtain

$$(3.66) \qquad \sum_{\substack{i=1 \\ i\neq k}}^{n+2} \mu_i s_i \sum_{j=1}^{n} y_j a_{ij} = 0 \ .$$

The above equation is true for any set of numbers y_j . In particular, then, it must be true for $y_j = x_j' - x_j$, $j=1,\ldots,n$, and thus we get

$$(3.67) \qquad \sum_{\substack{i=1 \\ i\neq k}}^{n+2} \mu_i s_i \sum_{j=1}^{n} a_{ij}(x_j'-x_j) = 0 \ .$$

If we sum (3.64) over all permissable indices of i , we get

$$(3.68) \qquad \sum_{\substack{i=1 \\ i\neq k}}^{n+2} \mu_i s_i \sum_{j=1}^{n} a_{ij}(x_j'-x_j) > 0 \ ,$$

which contradicts (3.67). Therefore, $\rho - \rho' < 0$ and thus $\rho' > \rho$ as was to be shown.

We can now state that the exchange algorithm converges in a finite number of steps. From the previous theorem, we know that the minimax error strictly increases during each cycle of the algorithm. Hence, the algorithm never returns to a previously considered subsystem. Since there are only a finite number of such subsystems, the algorithm must eventually reach the critical one whose existence is guaranteed by Theorem 3.3.

3.2.3.2 The Exchange Algorithm on a Computer

Our approach in this section will be to present the exchange algorithm step by step with explanations inserted where necessary.

From the last section, we know that we will always be considering some subset of $(n+1)$ equations. This can be most easily handled by an $(n+1)$ - component vector $I = \{i_1, i_2, \ldots, i_{n+1}\}$ which gives the indices of the $(n+1)$ equations presently under consideration. We suppose that the matrix A, the vector b, and a vector I have been given. The exchange algorithm then proceeds as follows:

1)
$$T_k \leftarrow (-1)^k \begin{vmatrix} a_{i_1,1} & \cdots & a_{i_1,n} \\ \vdots & & \vdots \\ a_{i_{k-1},1} & \cdots & a_{i_{k-1},n} \\ a_{i_{k+1},1} & \cdots & a_{i_{k+1},n} \\ \vdots & & \vdots \\ a_{i_{n+1},1} & \cdots & a_{i_{n+1},n} \end{vmatrix} , \quad k=1,\dots,n+1 .$$

Calculating the determinant is a standard problem in linear algebra. For this and all other problems in linear algebra we have adapted a set of programs given by Forsythe and Moler ([15], pp. 68-72). Their main computational tool is Gaussian elimination, with partial pivoting and iterative improvement. For completeness, these programs have been listed in the Appendix.

2) $s_k \leftarrow \operatorname{sgn} T_k$, $k=1,\dots,n+1$.

3)
$$\begin{bmatrix} d_{11} & \cdots & d_{1,n+1} \\ \vdots & & \vdots \\ d_{n+1,1} & \cdots & d_{n+1,n+1} \end{bmatrix} \leftarrow \begin{bmatrix} a_{i_1,1} & \cdots & a_{i_1,n} & s_1 \\ \vdots & & \vdots & \vdots \\ a_{i_{n+1},1} & \cdots & a_{i_{n+1},n} & s_{n+1} \end{bmatrix}^{-1}$$

4) $$\begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \rho \end{bmatrix} \leftarrow \begin{bmatrix} d_{11} & \cdots & d_{1,n+1} \\ \vdots & & \vdots \\ d_{n+1,1} & \cdots & d_{n+1,n+1} \end{bmatrix} \begin{bmatrix} b_{i_1} \\ \vdots \\ b_{i_{n+1}} \end{bmatrix}$$

The accumulation of inner products is carried out in double precision.

5) $$\rho_i \leftarrow b_i - \sum_{j=1}^{n} a_{ij} x_j \ , \quad i=1,\ldots,m \ .$$

6) $$|\rho_\alpha| = \max_{1 \le i \le m} |\rho_i| \ .$$

7) We test: Is $|\rho_\alpha| > \rho$? If the answer is no, then $[x_1,\ldots,x_n,\rho]$ gives the minimax solution and the associated minimax error. If the answer is yes, we have found it necessary to make a further test to see if $\alpha \in I$. If so, then we exit with the vector $[x_1,\ldots,x_n,\rho]$ as before. If not, then we proceed to the next step.

8) $u \leftarrow \operatorname{sgn} \rho_\alpha$.

A word of explanation is neçessary to explain the following steps in the algorithm. From the proof of Theorem 3.7, we see that we must express $s_\alpha A_\alpha$ as a linear combination of $s_1 A_{i_1}, \ldots, s_{n+1} A_{i_{n+1}}$. To do this, set

$$A_\alpha = \sum_{k=1}^{n+1} \lambda_k A_{i_k} . \tag{3.69}$$

The above equation can be written in matrix form

$$[\lambda_1, \ldots, \lambda_{n+1}] \begin{bmatrix} a_{i_1,1} & \cdots & a_{i_1,n} & s_1 \\ \vdots & & \vdots & \vdots \\ a_{i_{n+1},1} & \cdots & a_{i_{n+1},n} & s_{n+1} \end{bmatrix} \tag{3.70}$$

$$= [a_{\alpha 1}, \ldots, a_{\alpha n}, u] .$$

Therefore, the next step in the algorithm is given by

9) $[\lambda_1, \ldots, \lambda_{n+1}] = [a_{\alpha 1}, \ldots, a_{\alpha n}, u] \begin{bmatrix} d_{11} & \cdots & d_{1,n+1} \\ \vdots & & \vdots \\ d_{n+1,1} & \cdots & d_{n+1,n+1} \end{bmatrix}$

We must now determine k from (3.54). From the definition of an inverse, we have

$$\begin{bmatrix} d_{11} & \cdots & d_{1,n+1} \\ \vdots & & \vdots \\ d_{n+1,1} & \cdots & d_{n+1,n+1} \end{bmatrix} \begin{bmatrix} a_{i_1,1} & \cdots & a_{i_1,n} & s_1 \\ \vdots & & \vdots & \vdots \\ a_{i_{n+1},1} & \cdots & a_{i_{n+1},n} & s_{n+1} \end{bmatrix} =$$

$$(3.71) \qquad \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

and hence,

$$(3.72) \qquad \sum_{k=1}^{n+1} d_{n+1,k} A_{i_k} = 0 \; .$$

Comparing (3.72) with (3.46), we see that after the initial inverse has been calculated, the numbers $d_{n+1,k}$ may serve as the numbers T_k . From (3.69) we have

(3.73) $$u\,A_\alpha = \sum_{k=1}^{n+1} \lambda_k u s_k (s_k A_{i_k})$$

and therefore, the next step in the algorithm can be written

10) $$u\,\frac{\lambda_k}{d_{n+1,k}} = \max_{1\le i\le n+1} u\,\frac{\lambda_i}{d_{n+1,i}}\;.$$

The only thing left to do is to update the inverse. We are changing the k^{th} row of a coefficient matrix by replacing it by a known linear combination of rows, the coefficients being the λ_k of (3.69). The effect on the inverse is given by Handscomb ([18], pp. 79-80).

11) $$d_{ij} \leftarrow \begin{cases} d_{ij}/\lambda_j\;, & j=k;\ i=1,\ldots,n+1\;. \\ d_{ij} - \frac{\lambda_j}{\lambda_k}\,d_{ik}\;, & j\ne k;\ i=1,\ldots,n+1\;. \end{cases}$$

12) $i_k \leftarrow \alpha$.

The algorithm now returns to step 4.

A program to compute the minimax solution of (3.4) is given in the Appendix.

CHAPTER IV

HAAR'S CONDITION

4.1 An Example

In the previous chapter we presented a detailed exposition of the exchange algorithm for computing the minimax solution of an overdetermined, inconsistent system of linear equations $Ax = b$ (see Section 3.4). The theoretical basis of this algorithm relied heavily on the assumption that the row vectors comprising the coefficient matrix A satisfied Haar's condition (see Definition 3.1). To see the difficulties arising when Haar's condition is violated, consider the problem of determining the minimax solution of the following system of linear equations

$$\begin{aligned} 2x_1 + x_2 &= 6.9 \\ 3x_1 + x_2 &= 7.2 \\ x_1 + x_2 &= 3.0 \\ x_1 - x_2 &= 1.0 \\ 2x_1 + 4x_2 &= 11.1 \\ x_1 + 2x_2 &= 7.0 \end{aligned} \tag{4.1}$$

We note that Haar's condition is violated since the vectors (2,4) and (1,2) corresponding to the fifth and sixth equations are linearly dependent. At each stage, the exchange algorithm considers subsystems of three equations. As in Section 3.2.3.2, we will let I be a 3-component vector giving the indices of the equations presently under consideration. Initially, suppose $I = [1,2,3]$. We present the computations of the exchange algorithm given in Section 3.2.3.2.

Step 2

1) $T_1 = (-1) \begin{vmatrix} 3 & 1 \\ 1 & 1 \end{vmatrix} = -2,\ T_2 = 1,\ T_3 = 1$.

2) $s_1 = \operatorname{sgn} T_1 = -1,\ s_2 = 1,\ s_3 = 1$.

3)
$$D = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} & \frac{5}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} & \frac{5}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 6.9 \\ 7.2 \\ 3.0 \end{bmatrix} = \begin{bmatrix} 2.1 \\ 1.8 \\ -0.9 \end{bmatrix}$$

5) $\rho_1 = 0.9$, $\rho_2 = -0.9$, $\rho_3 = -0.9$,

$\rho_4 = 0.7$, $\rho_5 = -0.3$, $\rho_6 = 1.3$.

6) $|\rho_6| = \max\limits_{1 \le i \le 6} |\rho_i|$; $\alpha = 6$.

7) $|\rho_6| > \rho$. $6 \not\in I = [1,2,3]$.

8) $u = \operatorname{sgn} \rho_6 = 1$.

9) $$[\lambda_1, \lambda_2, \lambda_3] = [1,2,1] \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} & \frac{5}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

$$= [\tfrac{1}{2}, -\tfrac{3}{4}, \tfrac{9}{4}] .$$

10) $\dfrac{u\ \lambda_3}{d_{33}} = 9 = \max\limits_{1 \le i \le 3} u\ \dfrac{\lambda_i}{d_{3i}}$; $k = 3$.

11) $$D = \begin{bmatrix} \frac{1}{9} & \frac{1}{3} & -\frac{2}{9} \\ \frac{2}{9} & -\frac{1}{3} & \frac{5}{9} \\ -\frac{5}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix}$$

12) $I = [1,2,6]$.

<u>Step 2</u>

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} \frac{1}{9} & \frac{1}{3} & -\frac{2}{9} \\ \frac{2}{9} & -\frac{1}{3} & \frac{5}{9} \\ -\frac{5}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix} \begin{bmatrix} 6.9 \\ 7.2 \\ 7.0 \end{bmatrix} = \begin{bmatrix} \frac{29}{18} \\ \frac{136}{45} \\ -\frac{59}{90} \end{bmatrix}$$

5) $\rho_1 = \frac{59}{90}$, $\rho_2 = -\frac{59}{90}$, $\rho_3 = \frac{147}{90}$,

$\rho_4 = \frac{217}{90}$, $\rho_5 = -\frac{379}{90}$, $\rho_6 = -\frac{59}{90}$.

6) $|\rho_5| = \max\limits_{1 \le i \le 6} |\rho_i|$. $\alpha = 5$.

7) $|\rho_5| > \rho$. $5 \notin I$.

8) $u = \operatorname{sgn} \rho_5 = -1$.

9)

$$[\lambda_1, \lambda_2, \lambda_3] = [2, 4, -1] \begin{bmatrix} \frac{1}{9} & \frac{1}{3} & -\frac{2}{9} \\ \frac{2}{9} & -\frac{1}{3} & \frac{5}{9} \\ -\frac{5}{9} & \frac{1}{3} & \frac{1}{9} \end{bmatrix}$$

$$= [\tfrac{5}{3}, -1, \tfrac{5}{3}] \ .$$

10) $\dfrac{u\ \lambda_1}{d_{31}} = \dfrac{u\ \lambda_2}{d_{32}} = 3 = \max\limits_{1 \le i \le 3} u \dfrac{\lambda_i}{d_{3i}}$; $k = 1$ or $k = 2$.

We have shown in Chapter III the maximum of the above ratios would be unique if Haar's condition were satisfied. We are presented with two choices above, and we will consider both of them.

Case 1 (k = 2)

11)
$$D = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

12) $I = [1,5,6]$.

Step 3

4)
$$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} \end{bmatrix} \begin{bmatrix} 6.9 \\ 11.1 \\ 7.0 \end{bmatrix} = \begin{bmatrix} \frac{97}{30} \\ \frac{14}{10} \\ \frac{29}{30} \end{bmatrix}$$

5)
$$\rho_1 = -\frac{29}{30} , \quad \rho_2 = -\frac{39}{10} , \quad \rho_3 = -\frac{49}{30} ,$$
$$\rho_4 = -\frac{25}{30} , \quad \rho_5 = -\frac{29}{30} , \quad \rho_6 = \frac{29}{30} .$$

6) $|\rho_2| = \max_{1 \le i \le 6} |\rho_i|$; $\alpha = 2$.

7) $|\rho_2| > \rho$. $2 \not\in I$.

8) $u = \operatorname{sgn} \rho_2 = -1$.

9)

$$[\lambda_1,\lambda_2,\lambda_3] = [3,1,-1] \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$= [\tfrac{5}{3}, -\tfrac{1}{3}, \tfrac{1}{3}] \ .$$

10) $\dfrac{u\ \lambda_1}{d_{31}} = \dfrac{5/3}{0}$, $\dfrac{u\ \lambda_2}{d_{32}} = -3$, $\dfrac{u\ \lambda_3}{d_{33}} = -\ 1/2$.

Our computations break down at this point because of the division by zero. Let us therefore consider the second alternative at stage 10 in step 2.

Case 2 (k = 1)

11)

$$D = \begin{bmatrix} \frac{1}{15} & \frac{2}{5} & -\frac{1}{3} \\ \frac{2}{15} & -\frac{1}{5} & \frac{1}{3} \\ -\frac{1}{3} & 0 & \frac{2}{3} \end{bmatrix}$$

12) $I = [5,2,6]$.

We now carry out the computations in Step 3. However, the computations will again break down at stage 10, for in calculating $u\ \lambda_2/d_{32}$, we will once more encounter a zero divisor. Thus the exchange algorithm has failed in this example with initial choice of I given by $I = [1,2,3]$.

There are two possible ways of working out of this dilemma. First, we can try to show in the context of the Dirichlet problem that Haar's condition is automatically satisfied. If this approach fails, we can try to modify the exchange algorithm so that it will work even though Haar's condition is violated. Let us examine the first of these alternatives.

4.2 Unisolvent Functions

Consider a set of functions $g_1,\ldots,g_n$ defined on a point set X. We then have

Definition 4.1. *The set of functions* $g_1,\ldots,g_n$ *is said to be unisolvent if the determinant*

$$(4.2) \qquad G = \begin{vmatrix} g_1(x_1) & \cdots & g_n(x_1) \\ \vdots & & \vdots \\ g_1(x_n) & \cdots & g_n(x_n) \end{vmatrix}$$

is non-zero for any n *distinct points* $x_1,\ldots,x_n$ *in* X .

Suppose now that we have a set of unisolvent functions $g_1,\ldots,g_n$ on a point set X. Then the coefficient matrix of the linear system

$$(4.3) \qquad \sum_{j=1}^{n} c_j g_j(x_i) = f(x_i) \ , \quad i=1,\ldots,m\geq n$$

obviously satisfies Haar's condition.

In the context of Dirichlet's problem, the linear system is

$$(4.4) \qquad \sum_{j=1}^{n} c_j g_j(x_i,y_i) = f(x_i,y_i) \ , \quad i=1,\ldots,m>n \ ,$$

where the m points are selected from the contour C. As a consequence of the following assertion, we should always approximate f by at least three functions.

Theorem 4.1. *No set of two functions* g_1,g_2 *are unisolvent on a simple closed curve* C .

Proof: Suppose the determinant

$$(4.5) \qquad G = \begin{vmatrix} g_1(x_1,y_1) & g_2(x_1,y_1) \\ g_1(x_2,y_2) & g_2(x_2,y_2) \end{vmatrix}$$

is non-zero for a certain pair of points (x_1,y_1) and (x_2,y_2) . By a continuous motion around the curve C , these two points can be interchanged without becoming coincident. In effect, we have then interchanged two rows in the determinant in (4.5). Hence G has changed sign. Therefore, at some intermediate position the determinant must vanish, which was to be shown.

Let us now turn to the case where $g_1,\ldots,g_n$ are the harmonic polynomials. We must then examine the behavior

of the following determinant:

$$(4.6) \qquad G_n = \begin{vmatrix} 1 & \mathrm{Re}\, z_1 & \mathrm{Im}\, z_1 & \cdots & \mathrm{Re}\, z_1^n & \mathrm{Im}\, z_1^n \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \mathrm{Re}\, z_{2n+1} & \mathrm{Im}\, z_{2n+1} & \cdots & \mathrm{Re}\, z_{2n+1}^n & \mathrm{Im}\, z_{2n+1}^n \end{vmatrix}$$

where $z_1, \ldots, z_{2n+1}$ are distinct points from C . To show a simple example of the vanishing of the above determinant, suppose $n = 1$. Then (4.6) reduces to

$$(4.7) \qquad G_1 = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}$$

where $z_k = x_k + iy_k$, $k=1,2,3$. $G_1 = 0$ if the points (x_1,y_1), (x_2,y_2), (x_3,y_3) lie on any straight line. Therefore, if C contains a straight line segment, the functions 1, Re z, Im z are not unisolvent on C .

The condition for the vanishing of G_n is that there exist some linear combination of the functions 1, Re z, Im z, ..., Re z^n, Im z^n , with coefficients not all zero, which vanishes at the points $z_1, \ldots, z_{2n+1}$. Therefore, the condition for the nonvanishing of G_n is

that the points $z_1,\dots,z_{2n+1}$ do not lie on any curve of the form $H_n(z) = 0$, where $H_n(z)$ is a linear combination of the first $2n+1$ harmonic polynomials. If we wish this to be true for any $2n+1$ points on C , then we must have that C does not intersect any curve of the form $H_n(z) = 0$ at more than $2n$ points.

For an arbitrary simple closed curve C , this is a difficult condition to check. Now $H_n(z) = 0$ is an algebraic curve of degree n . An algebraic curve of degree n can intersect an algebraic curve of degree 2 in at most $2n$ distinct points. Hence, if C is an ellipse or a circle, the harmonic polynomials are unisolvent on C .

A more extensive discussion of these matters can be found in Curtiss [9,10].

4.3 Cheney's Perturbation Method

The results of the previous section are rather discouraging in that the only interesting boundaries on which we can be sure that the harmonic polynomials are unisolvent are circles and ellipses. However, an alternative way of handling the problem is a perturbation technique suggested by Cheney ([13], p. 51). The method is suggested by the following assertion.

Theorem 4.2. *If for a square matrix* $A = (a_{ij})$, $|A| = 0$, *then for sufficiently small* $\varepsilon \neq 0$, $|A-\varepsilon I| \neq 0$.

Proof: Suppose that $|A - \lambda I| = 0$. Then λ is an eigenvalue of the matrix (a_{ij}) , $i,j=1,\ldots,n$. There are at most n such eigenvalues, say $\lambda_1,\ldots,\lambda_n$. If ε is chosen such that $0 < |\varepsilon| < \max\limits_{1 \le j \le n} |\lambda_j|$, $\lambda_j \ne 0$, then the theorem is established.

To show how this method works, we again consider the problem of finding a minimax solution of (4.1). Since the vectors comprising the fifth and sixth equations are dependent, Theorem 4.1 suggests that we first determine the minimax solution to

$$
(4.9) \qquad
\begin{aligned}
2x_1 + x_2 &= 6.9 \\
3x_1 + x_2 &= 7.2 \\
x_1 + x_2 &= 3.0 \\
x_1 - x_2 &= 1.0 \\
(2+\varepsilon)x_1 + 4x_2 &= 11.1 \\
x_1 + (2+\varepsilon)x_2 &= 7.0 ,
\end{aligned}
$$

where $0 < |\varepsilon| < 4$. We will follow the same format as in Section 4.1. Initially, we again choose $I = [1,2,3]$.

Step 1

1) $T_1 = -2$, $T_2 = 1$, $T_3 = 1$.

2) $s_1 = -1$, $s_2 = 1$, $s_3 = 1$.

3)
$$D = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{3}{4} & \frac{5}{4} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} .$$

4)
$$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = D \begin{bmatrix} 6.9 \\ 7.2 \\ 3.0 \end{bmatrix} = \begin{bmatrix} 2.1 \\ 1.8 \\ -0.9 \end{bmatrix} .$$

5) $\rho_1 = 0.9$, $\rho_2 = -0.9$, $\rho_3 = -0.9$,

$\rho_4 = 0.7$, $\rho_5 \sim -0.3$, $\rho_6 \sim 1.3$.

In determining ρ_5 and ρ_6 , we have assumed that ε is

insignificantly small compared to 0.3. We have replaced the equality sign by the symbol "~" whenever a similar process has been carried out.

6) $|\rho_6| = \max\limits_{1\leq i\leq 6} |\rho_i|$; $\alpha = 6$.

7) $|\rho_6| > \rho$. $6 \not\in I$.

8) $u = 1$.

9) $[\lambda_1,\lambda_2,\lambda_3] = [1,2+\varepsilon,1]\ D = [\frac{1}{2}(1+\varepsilon),\ -\frac{3}{4}(1+\varepsilon),\ \frac{1}{4}(9+5\varepsilon)]$.

10) $\dfrac{u\ \lambda_3}{d_{33}} \sim 9 = \max\limits_{1\leq i\leq 3} u\ \dfrac{\lambda_i}{d_{3i}}$; $k = 3$.

11) $$D = \frac{1}{9+5\varepsilon}\begin{bmatrix} 1+\varepsilon & 3+\varepsilon & -2 \\ 3 & -3 & 5 \\ -5-3\varepsilon & 3+2\varepsilon & 1 \end{bmatrix} .$$

12) $I = [1,2,6]$.

Step 2

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \frac{1}{10(9+5\varepsilon)} \begin{bmatrix} 145 + 141\varepsilon \\ 272 \\ -59 - 63\varepsilon \end{bmatrix}$$

5) $\rho_1 \sim \frac{59}{90}$, $\rho_2 \sim -\frac{59}{90}$, $\rho_3 \sim -\frac{147}{90}$,

$\rho_4 \sim \frac{217}{90}$, $\rho_5 \sim -\frac{379}{90}$, $\rho_6 \sim -\frac{59}{90}$.

6) $|\rho_5| = \max_{1 \le i \le 6} |\rho_i|$; $\alpha = 5$.

7) $|\rho_5| > \rho$. $5 \notin I$.

8) $u = -1$.

9) $[\lambda_1,\lambda_2,\lambda_3] = [2+\varepsilon,4,-1]\ D$

$$= \frac{1}{9+5\varepsilon}\ [15+6\varepsilon+\varepsilon^2,\ -9+3\varepsilon+\varepsilon^2,\ 15-2\varepsilon]\ .$$

10) $\frac{u\ \lambda_1}{d_{31}} = \frac{15+6\varepsilon+\varepsilon^2}{5+3\varepsilon}\ ;\ \frac{u\ \lambda_2}{d_{32}} = \frac{9-3\varepsilon-\varepsilon^2}{3+2\varepsilon}\ ;\ \frac{u\ \lambda_3}{d_{33}} = -15+2\varepsilon\ .$

For $\varepsilon > 0$ and sufficiently small, the maximum of the above ratios is $\frac{u\ \lambda_1}{d_{31}}$. Therefore, $k = 1$.

11) $D = \frac{1}{(9+5\varepsilon)(15+6\varepsilon+\varepsilon^2)}\ \times$

$$\begin{bmatrix} (1+\varepsilon)(9+5\varepsilon) & 54+39\varepsilon+5\varepsilon^2 & -45-25\varepsilon \\ 2(9+5\varepsilon) & -27-24\varepsilon-5\varepsilon^2 & 45+34\varepsilon+5\varepsilon^2 \\ (-5-3\varepsilon)(9+5\varepsilon) & 36\varepsilon+29\varepsilon^2+5\varepsilon^2 & 90+41\varepsilon-5\varepsilon^2 \end{bmatrix}$$

12) $I = [5,2,6]$.

Step 3

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \frac{1}{10(9+5\varepsilon)(15+6\varepsilon+\varepsilon^2)} \begin{bmatrix} 1737 + 2612\varepsilon + 915\varepsilon^2 \\ 3204 + 1762\varepsilon - 10\varepsilon^2 \\ 1305 - 310\varepsilon - 7\varepsilon^2 + 360\varepsilon^3 \end{bmatrix}$$

5) $\rho_1 \sim \frac{2637}{1350}$, $\rho_2 \sim \frac{1305}{1350}$, $\rho_3 \sim -\frac{891}{1350}$,

$\rho_4 \sim \frac{2817}{1350}$, $\rho_5 \sim -\frac{1305}{1350}$, $\rho_6 \sim \frac{1305}{1350}$.

6) $|\rho_4| = \max_{1 \le i \le 6} |\rho_i|$; $\alpha = 4$.

7) $|\rho_4| > \rho$. $4 \notin I$.

8) $u = 1$.

9) $[\lambda_1,\lambda_2,\lambda_3] = [1,-1,1]\ D$

$$= \frac{1}{(9+5\varepsilon)(15+6\varepsilon+\varepsilon^2)} \times$$

$$[(-6-2\varepsilon)(9+5\varepsilon),\ 81+99\varepsilon+39\varepsilon^2+5\varepsilon^3, -18\varepsilon-10\varepsilon^2].$$

10) $\dfrac{u\ \lambda_2}{d_{32}} = \max\limits_{1\le i\le 3} u\ \dfrac{\lambda_i}{d_{3i}}\ ;\quad k = 2\ .$

11) $$D = \begin{bmatrix} \frac{1}{3+\varepsilon} & \frac{6+\varepsilon}{(3+\varepsilon)^2} & -\frac{3}{(3+\varepsilon)^2} \\ 0 & -\frac{1}{3+\varepsilon} & \frac{1}{3+\varepsilon} \\ -\frac{1}{3+\varepsilon} & \frac{4\varepsilon+\varepsilon^2}{(3+\varepsilon)^2} & \frac{6+\varepsilon}{(3+\varepsilon)^2} \end{bmatrix}$$

12) $I = [5,4,6]\ .$

Step 4

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} \frac{183+121\varepsilon}{10(3+\varepsilon)^2} \\ \frac{6}{3+\varepsilon} \\ \frac{87-\varepsilon+10\varepsilon^2}{10(3+\varepsilon)^2} \end{bmatrix}$$

5) $\rho_1 \sim \frac{75}{90}$, $\rho_2 \sim -\frac{81}{90}$, $\rho_3 \sim -\frac{93}{90}$,

$\rho_4 \sim \frac{87}{90}$, $\rho_5 \sim -\frac{87}{90}$, $\rho_6 \sim \frac{87}{90}$.

6) $|\rho_3| = \max_{1 \le i \le 6} |\rho_i|$; $\alpha = 3$.

7) $|\rho_3| > \rho$. $3 \notin I$.

8) $u = -1$.

9) $[\lambda_1,\lambda_2,\lambda_3] = [1,1,-1]\ D$

$$= [\ \frac{2}{3+\varepsilon}\ ,\ \frac{3-4\varepsilon-\varepsilon^2}{(3+\varepsilon)^2}\ ,\ -\frac{6}{(3+\varepsilon)^2}\ ]\ .$$

10) $2 \sim \frac{u\ \lambda_1}{d_{31}} = \max_{1\le i\le 3} u \frac{\lambda_i}{d_{3i}}\ ;\quad k = 1\ .$

11) $D = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & -\frac{1}{3+\varepsilon} & \frac{1}{3+\varepsilon} \\ -\frac{1}{2} & \frac{1+\varepsilon}{6+2\varepsilon} & \frac{1}{3+\varepsilon} \end{bmatrix}$

12) $I = [3,4,6]\ .$

Step 5

4) $\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{6}{3+\varepsilon} \\ \frac{6-2\varepsilon}{6+2\varepsilon} \end{bmatrix}$

5) $\rho_1 \sim 0.9$, $\rho_2 \sim -0.8$, $\rho_3 \sim -1.0$,
$\rho_4 \sim 1.0$, $\rho_5 \sim -0.9$, $\rho_6 \sim 1.0$.

6) $|\rho_3| = |\rho_4| = |\rho_6| = \max\limits_{1 \leq i \leq 6} |\rho_i|$; $\alpha = 3$ or 4 or 6 .

7) $|\rho_3| = |\rho_4| = |\rho_6| > \rho$; $\alpha \in I$.

The last line shows that the critical subsystem has been reached. To find the minimax solution of (4.1) we let $\varepsilon \to 0$ to obtain $x_1 = x_2 = 2$ and $\rho = 1$.

The technique shown above is not convenient for use on a computer since ε has been left as a variable in the computations. The same method employed on a computer would require the use of symbol manipulation, which is not readily available. The only recourse would seem to be to introduce a sequence of ε's say $\varepsilon_1 > \varepsilon_2 > \ldots \varepsilon_n \to 0$ and examine what happens. If no significant changes are observed after a certain ε_N , then one could accept the minimax solution to the problem with ε_N appearing as a good approximation to the minimax solution of the original problem.

The problem of finding the minimax solution of (4.1) was attacked using the computer program MINMAX given in the Appendix. This program assumes that Haar's condition is satisfied. The correct solution was obtained using both single and double precision arithmetic, using as initial guess $I = [1,2,3]$. The computations did not break down (as in the exact calculation shown in Section 4.1) because roundoff errors introduced during computations yielded very small numbers rather than zeros at the critical stages. In effect, the roundoff errors perturbed the original system, which did not satisfy Haar's condition, into an adjacent system that did. We present some of the results of the single precision computations to 3 significant digits. (The initial subsystem is given by $I = [1,2,3]$).

Step 2

1) $T_1 = -2$, $T_2 = 1$, $T_3 = 1$.

2) $s_1 = -1$, $s_2 = 1$, $s_3 = 1$.

3) $$D = \begin{bmatrix} 0.0 & 0.5 & -0.5 \\ 0.5 & -0.75 & 1.25 \\ -0.5 & 0.25 & 0.25 \end{bmatrix}$$

4) $$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} 2.10 \\ 1.80 \\ -0.90 \end{bmatrix}$$

5) $\rho_1 = 0.9$, $\rho_2 = -0.9$, $\rho_3 = -0.9$,

$\rho_4 = 0.7$, $\rho_5 = -0.3$, $\rho_6 = 1.30$.

6) $|\rho_6| = \max\limits_{1\le i\le 6} |\rho_i|$; $\alpha = 6$.

7) $|\rho_6| > \rho$. $6 \not\in I$.

8) $u = 1$.

9) $[\lambda_1,\lambda_2,\lambda_3] = [0.5,-0.75,2.25]$.

10) $9.0 = \frac{u\ \lambda_3}{d_{33}} = \max_{1\le i\le 3} u \frac{\lambda_i}{d_{3i}}$; $k = 3$.

11)
$$D = \begin{bmatrix} 0.111 & 0.333 & -0.222 \\ 0.222 & -0.333 & 0.556 \\ -0.556 & 0.333 & 0.111 \end{bmatrix}$$

12) $I = [1,2,6]$.

Step 2

4)
$$\begin{bmatrix} x_1 \\ x_2 \\ \rho \end{bmatrix} = \begin{bmatrix} 1.61 \\ 3.02 \\ -0.656 \end{bmatrix}$$

5) $\rho_1 = 0.656$, $\rho_2 = -0.656$, $\rho_3 = -1.63$,

$\rho_4 = 2.41$, $\rho_5 = -4.21$, $\rho_6 = -0.656$.

6) $|\rho_5| = \max_{1 \le i \le 6} |\rho_i|$; $\alpha = 5$.

7) $|\rho_5| > \rho$. $5 \notin I$.

8) $u = -1$.

9) $[\lambda_1, \lambda_2, \lambda_3] = [1.67, -1.0, 1.67]$.

10) $3.0 = \dfrac{u\ \lambda_2}{d_{32}} = \max_{1 \le i \le 3} u \dfrac{\lambda_i}{d_{3i}}$; $k = 2$.

11)
$$D = \begin{bmatrix} 0.667 & -0.333 & 0.333 \\ -0.333 & 0.333 & 0.212\times10^{-6} \\ -0.212\times10^{-6} & -0.333 & 0.667 \end{bmatrix}$$

12) $I = [1,5,6]$.

Note in stage (11) that $d_{31} \neq 0$. In the exact computation, $d_{31} = 0$. The computations therefore do not break down. The algorithm then considers successively the subsystems [2,5,6], [3,5,6], and [3,4,6]. It reports that the last subsystem is the critical one and gives the minimax solution as $x_1 = x_2 = 2.0$ and the associated minimax error as $\rho = 1.0$.

This result is somewhat encouraging. It should be noted, however, that the reverse process is also possible: that is, it is possible for roundoff errors to perturb a system satisfying Haar's condition into an adjacent one that violates Haar's condition. Individual cases will probably require individual treatment.

CHAPTER V

MINIMAX APPROXIMATIONS ON A CONTINUUM

5.1 Introduction

In the previous two chapters, we solved the problem of determining a vector $c = (c_1,\ldots,c_n)$ to minimize

$$(5.1) \qquad M_1 = \max_{1 \le i \le m} \left| f(x_i,y_i) - \sum_{j=1}^{n} c_j g_j(x_i,y_i) \right| ,$$

where (x_i,y_i), $i=1,\ldots,m$, were distinct points on the contour C. Intuitively, we feel that if the number of points becomes large, leaving no wide gaps along C, the minimax error associated with minimizing (5.1) should be a good approximation to the minimax error associated with minimizing

$$(5.2) \qquad M_2 = \max_{(x,y) \varepsilon C} \left| f(x,y) - \sum_{j=1}^{n} c_j g_j(x,y) \right| .$$

Thus, one approach to finding the minimax error on the curve C would be as follows: Compute the minimax error for m points. Double the number of points and again

calculate the minimax error. If the minimax error does not change significantly, then stop. Otherwise, once more double the number of points and continue as before.

A more elegant approach was first conceived by Remes in 1934. The statement of the algorithm and the proof of its convergence is from Cheney ([3], pp. 95-96).

5.2 Remes Algorithm

This algorithm only requires that the functions $f, g_1, \ldots, g_n$ be continuous on the curve C. We wish to determine a vector $c = (c_1, \ldots, c_n)$ for which the deviation

$$M(c) = \max_{s \varepsilon C} \left| f(s) - \sum_{j=1}^{n} c_j g_j(s) \right| \tag{5.3}$$

is a minimum. We may assume without loss of generality that $g_1, \ldots, g_n$ are linearly independent, for if they are not, we can replace them by a smaller number of functions that are independent without raising the minimum deviation

$$\rho = \min_{c} M(c) \ . \tag{5.4}$$

We define a residual function $r(c,s)$ by

$$(5.5) \qquad r(c,s) = f(s) - \sum_{j=1}^{n} c_j g_j(s) \ .$$

Then Remes algorithm is as follows: At the k^{th} step we are given a finite subset S^k of C . Select a coefficient vector c^k to minimize the function

$$(5.6) \qquad M^k(c) = \max_{s \varepsilon S^k} |r(c,s)| \ .$$

This can be accomplished by the procedure given in Chapter III. Choose $s^k \varepsilon C$ to maximize $|r(c^k,s)|$. Thus,

$$(5.7) \qquad |r(c^k,s^k)| = M(c^k) \ .$$

Now start again with the finite set $S^{k+1} = S^k \cup \{s^k\}$. At the beginning, S^1 may be arbitrary except that the set of n-tuples $\hat{s} = [g_1(s),\ldots,g_n(s)]$ corresponding to $s \varepsilon S^1$ should be of rank n . The algorithm stops when $|r(c^k,s^k)|$ is insignificantly different from $\min_{c} \max_{s \varepsilon S^k} |r(c,s)|$. A theorem associated with Remes algorithm is

<u>Theorem 5.1</u>. *$\lim_{k \to \infty} M^k(c^k) = \rho$. The sequence c^k is bounded and its cluster points minimize M .*

Proof: We define

$$\|c\|_1 = \sum_{j=1}^{n} |c_j| \;. \tag{5.8}$$

It can be shown that this function defines a norm in the linear space R_n. By our assumption on S^1, it follows that

$$\phi = \min_{\|c\|_1=1} \; \max_{s\varepsilon S^1} \left| \sum_{j=1}^{n} c_j g_j(s) \right| > 0 \;. \tag{5.9}$$

Consequently,

$$\begin{aligned} M^1(c) &= \max_{s\varepsilon S^1} \left|f(s) - \sum_{j=1}^{n} c_j g_j(s)\right| \\ &= \left\|f - \sum_{j=1}^{n} c_j g_j\right\| \\ &\geq \left\|\sum_{j=1}^{n} c_j g_j\right\| - \|f\| \\ &= \max_{s\varepsilon S^1} \left|\sum_{j=1}^{n} c_j g_j(s)\right| - \max_{s\varepsilon S^1} |f(s)| \;. \end{aligned} \tag{5.10}$$

Now let

$$\|c\|_1 = K \; . \tag{5.11}$$

Then

$$\max_{s\varepsilon S^1} \left|\sum_{j=1}^{n} c_j g_j(s)\right| = K \max_{s\varepsilon S^1} \left|\sum_{j=1}^{n} c_j/K \; g_j(s)\right| \; . \tag{5.12}$$

Therefore, we may rewrite (5.10) as

$$M^1(c) \geq \|c\|_1 \max_{s\varepsilon S^1} \left|\sum_{j=1}^{n} c_j g_j(s)\right| - \max_{s\varepsilon S^1} |f(s)| \; , \tag{5.13}$$

and hence by (5.9),

$$M^1(c) \geq \|c\|_1 \, \phi - \max_{s\varepsilon S^1} |f(s)| \; . \tag{5.14}$$

Now suppose that $\|c\|_1 > 2 \max_{s\varepsilon S^1} |f(s)|/\phi$. Then

$$M^k(c) \geq M^1(c)$$

$$\text{(5.15)} \qquad > \max_{s \in S^1} |f(s)|$$

$$\geq M^k(\Theta) \ ,$$

where Θ is the zero vector. Therefore, c does not enter the competition to minimize any of the functions M^k . Thus the sequence c^k generated by the algorithm is bounded. Now we have

$$\text{(5.16)} \qquad S^k \subset S^{k+1} \subset C \ ,$$

and hence

$$\text{(5.17)} \qquad M^k(c) \leq M^{k+1}(c) \leq M(c)$$

for any vector c . Therefore,

$$\text{(5.18)} \qquad M^k(c^k) \leq M^{k+1}(c^{k+1}) \leq \rho \ .$$

Thus for some $\varepsilon \geq 0$, $M^k(c^k) \rightarrow \rho - \varepsilon$. We must show that $\varepsilon = 0$. Since

$$(5.19) \qquad |r(b,s) - r(c,s)| = |\sum_{j=1}^{n} (c_j - b_j)\, g_j(s)|$$

$$\leq T \, \|b-c\|_1 ,$$

where

$$(5.20) \qquad T = \max_{1 \leq j \leq n} \; \max_{s \varepsilon C} |g_j(s)| ,$$

it follows that

$$(5.21) \qquad |r(b,s)| \leq |r(c,s)| + T \, \|b-c\|_1$$

and furthermore

$$(5.22) \qquad M(b) \leq M(c) + T \, \|b-c\|_1 .$$

Suppose now that $\varepsilon > 0$. Let b denote any cluster point

of the sequence c^k . For any $\delta > 0$, we may find an index k such that

$$\|b-c^k\|_1 < \delta \tag{5.23}$$

and an index $i > k$ such that

$$\|b-c^i\|_1 < \delta . \tag{5.24}$$

Then

$$\|c^i-c^k\|_1 \le 2\delta \tag{5.25}$$

and

$$\begin{aligned} \rho \le M(b) &\le M(c^k) + T\delta \\ &= |r(c^k,s^k)| + T\delta \\ &\le |r(c^i,s^k)| + 3T\delta \qquad (5.26) \\ &\le M^i(c^i) + 3T\delta \\ &\le \rho - \varepsilon + 3T\delta . \end{aligned}$$

Picking δ small enough so that $3T\delta < \varepsilon$ yields a contradiction. Therefore, $\varepsilon = 0$ and (5.26) can be rewritten

$$\rho \leq M(b) \leq \rho + 3T\delta \ . \tag{5.27}$$

Letting $\delta \to 0$ proves the assertion.

Remes algorithm requires that we locate the extremum of the residual function $r(c,s)$. This can be a difficult problem even in well-behaved cases. There would be no problem in constructing an example to defeat any particular method.

We have chosen the obvious and simple method of calculating the residual on a discrete set of points chosen in advance. Originally, this discrete set was chosen as a set of equally spaced points. Examination of the residual function, however, has shown that it has approximately the form of a Chebyshev polynomial ([13], p. 60), with the extrema crowded towards the ends of the range under consideration. Therefore, the discrete point set was chosen to be the zeros of a Chebyshev polynomial. As a rule of thumb, if we are approximating f by n functions $g_1,\ldots,g_n$, then the discrete point set was chosen to be the zeros of the (20n)th Chebyshev polynomial, although no significant change in the minimax error was evident if

we took the discrete point set to be the zeros of the Chebyshev polynomial of (10n)th degree or higher.

CHAPTER VI

NUMERICAL RESULTS

The present experiments were performed using the IBM System 360 FORTRAN IV source language and the IBM 360/67 computer system at the University of Alberta, Edmonton.

6.1 Test Problems and Results

For clarity, we shall elaborate the meaning of several terms which occur in the tabulation of the data.

Boundary C: The boundary of the region of space over which the solution is sought.

Boundary values: A continuous function f defined on the boundary. If there are more than one, then the functions will be labelled (a), (b), ...

Approximating functions: Harmonic functions selected from a set of harmonic functions which is complete over the space of functions which are continuous on the boundary. Usually, these functions will be selected from the set of harmonic polynomials (see Section 2.3).

Boundary-value problem: This will be either the Dirichlet problem or the mixed problem for Laplace's equation (see [16], p. 169, for the definition of the mixed problem).

Minimax error, ρ_n : The computed minimax error when n approximating functions are employed.

Minimax solution: This will be written in the form $u \sim c_1 g_1 + \ldots + c_n g_n$, where u is the required solution of the Dirichlet problem, $g_1, \ldots, g_n$ are the approximating functions, and $c_1, \ldots, c_n$ are the computed coefficients that minimize the maximum error on the boundary.

Least-Squares solution: This will be of the same form as the minimax solution except that the coefficients minimize the sum of the squares of the errors at discrete points along the boundary.

Error at isolated interior point(s): The absolute value of the difference between the theoretical value and the computed value at the given point(s). This can be given only if the true solution is known.

6.1.1 Problem 1 For complete details see Davis [12].

Boundary: A bean-shaped region was obtained from a free-hand drawing on coordinate paper (see Figure 6.1). The boundary is "defined" by means of 43 points (see Table 6.1). These points are not distributed equally on the boundary; more points appear where the curvature is greater.

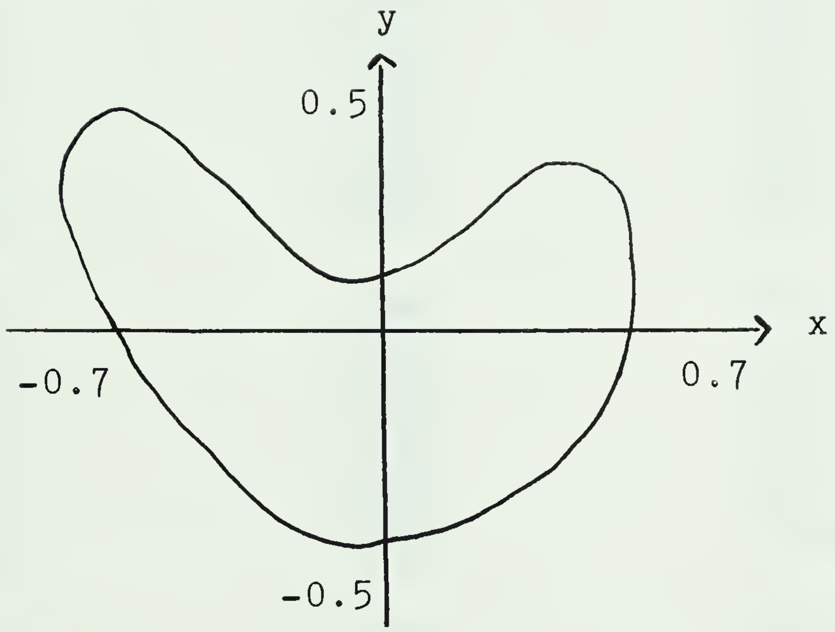


Figure 6.1 Bean-shaped region of Problem 1

[illegible]

Table 6.1 Points defining the bean-shaped region

Pt. no.	Abscissa	Ordinate
1	.000	.110
2	-.050	.108
3	-.100	.115
4	-.160	.150
5	-.220	.205
6	-.320	.300
7	-.400	.358
8	-.500	.420
9	-.550	.436
10	-.600	.430
11	-.644	.400
12	-.660	.350
13	-.655	.300
14	-.635	.200
15	-.595	.100
16	-.552	.000
17	-.500	-.105
18	-.440	-.200
19	-.400	-.250
20	-.350	-.300
21	-.300	-.344
22	-.204	-.400
23	-.100	-.436
24	.000	-.448
25	.100	-.442

Table 6.1 (Continued)

Pt. no.	Abscissa	Ordinate
26	.230	-.400
27	.300	-.350
28	.353	-.300
29	.430	-.200
30	.477	-.100
31	.510	.000
32	.522	.100
33	.520	.160
34	.500	.240
35	.456	.300
36	.400	.330
37	.360	.337
38	.300	.320
39	.250	.290
40	.200	.245
41	.150	.200
42	.100	.160
43	.050	.128

Boundary values:

(a) $f = 0, x \leq 0; \quad f = x, x > 0$.

(b) $f = 0, x \leq 0; \quad f = x^2, x > 0$.

(c) $f = 0, x \leq 0; \quad f = x^3, x > 0$.

(d) $f = 0, x \leq 0; \quad f = x^4, x > 0$.

(e) $f = 0, x \leq 0; \quad f = x^5, x > 0$.

(f) $f = x^2 + y^2$ (Torsion problem).

(g) $f = \exp(x^2 - xy + 2y^2)$.

(h) $f = \ln[(x + 1.5)^2 + y^2]$.

(i) $f = \ln[x + y + (y-1)^2]$.

(j) $f = e^x \cos y + \ln[x^2 + (y-1)^2]$.

Approximating functions: 1, $\mathrm{Re}\, z^n$, $\mathrm{Im}\, z^n$; $n=1,2,\ldots$.

Boundary value problem: The Dirichlet problem.

Minimax error: See Table 6.2.

<u>Minimax solution</u>: This is for function (j) for n = 5 .

$$\begin{aligned} u \sim 1.00244 &+ .997007 \text{ Re } z - 1.99034 \text{ Im } z \\ &+ 1.47952 \text{ Re } z^2 - .0110482 \text{ Im } z^2 \\ &+ .187437 \text{ Re } z^3 + .628927 \text{ Im } z^3 \\ &- .345989 \text{ Re } z^4 + .0152750 \text{ Im } z^4 \\ &- .133709 \text{ Re } z^5 - .322899 \text{ Im } z^5 \end{aligned}$$

<u>Least squares solution</u>: The coefficients are quoted from [12]. This is for the same problem as above.

$$\begin{aligned} u \sim 1.00173 &+ .997339 \text{ Re } z - 1.99119 \text{ Im } z \\ &+ 1.48065 \text{ Re } z^2 - .00949996 \text{ Im } z^2 \\ &+ .188957 \text{ Re } z^3 + .623677 \text{ Im } z^3 \\ &- .355600 \text{ Re } z^4 + .024526 \text{ Im } z^4 \\ &- .11960 \text{ Re } z^5 - .28034 \text{ Im } z^5 \end{aligned}$$

For this approximation the maximum error on the boundary is .0069.

Table 6.2 Results for bean-shaped region

	Minimax error ρ_n for function				
n	(a)	(b)	(c)	(d)	(e)
1	.142	.0823	.0477	.0266	.0145
2	.0870	.0410	.0232	.0135	.00776
3	.0725	.0284	.0131	.00657	.00357
4	.0613	.0234	.00961	.00378	.00156
5	.0562	.0182	.00757	.00321	.00138
6	.0480	.0164	.00662	.00276	.00123
7	.0446	.0141	.00546	.00233	.00105
8	.0414	.0126	.00483	.00202	.000909
9	.0382	.0113	.00416	.00174	.000783

	Minimax error ρ_n for function				
n	(f)	(g)	(h)	(i)	(j)
1	.190	.544	.613	.384	.277
2	.148	.338	.0322	.192	.0858
3	.0878	.184	.00930	.0940	.0258
4	.0707	.116	.00250	.0429	.0110
5	.0599	.0964	.000759	.0216	.00434
6	.0505	.0810	.000226	.0122	.00179
7	.0433	.0698	.0000707	.00879	.000738
8	.0371	.0607	.0000204	.00645	.000321
9	.0298	.0487	.000000545	.00491	.000134

Error at isolated interior points: The results (see Table 6.3) pertain to function (j) for both the minimax and least-squares solution.

Table 6.3 Discrepancies at points interior to bean along $y = 0$

x	minimax discrepancies	least-squares discrepancies
-.5	.0033	.0014
-.4	.0021	.0010
-.3	.0020	.0011
-.2	.0023	.0015
-.1	.0025	.0017
0	.0024	.0017
.1	.0020	.0013
.2	.0013	.0007
.3	.0006	.0001
.4	.0005	.0009

6.1.2 Problem 2

Boundary: The ellipse $x^2 + 4y^2 = 1$.

Boundary values:

(a) $f = 0,\ x \le 0;\quad f = x,\ x > 0$.

(b) $f = 0,\ x \le 0;\quad f = x^2,\ x > 0$.

(c) $f = 0,\ x \le 0;\quad f = x^3,\ x > 0$.

(d) $f = \ln[(x + 1.5)^2 + y^2]$.

Approximating functions: 1, Re z^n; $n=1,2,\ldots$. Only the real parts of z^n contribute since the boundary and boundary functions are symmetric with respect to the x axis. Thus only the top half of the boundary need be considered.

Boundary-value problem: The Dirichlet problem.

Minimax error: See Table 6.4.

Table 6.4 Results for the ellipse

	Minimax error ρ_n for function			
n	(a)	(b)	(c)	(d)
1	.250	.281	.317	.385
2	.0577	.0856	.134	.139
3	.0525	.0138	.0370	.0577
4	.0311	.0138	.00442	.0258
5	.0308	.0531	.00444	.0120
6	.0210	.0531	.00133	.00571
7	.0209	.00277	.00133	.00278
8	.0159	.00277	.000566	.00137

Exact value of solution for function (d) at the point (0,0) : $u(0,0) = \ln 2.25 = 0.8109$.

Computed value of approximating solution at the point (0,0) for n approximating functions: See Table 6.5.

Table 6.5 Computed solution for function (d) at (0,0)

n	computed value at (0,0)	discrepancy between computed and true solution
2	.6105	.2004
3	.8544	.0435
4	.8207	.0098
5	.8077	.0032
6	.8103	.0006
7	.8112	.0003
8	.8110	.0001
9	.8109	.0000

6.1.3 Problem 3

Boundary: Square given by the lines $x = 0$, $x = 1$, $y = 0$, $y = 1$.

Boundary values: (a)

$$f = \begin{cases} 1, & y = 1,\ 0 < x < 1 . \\ 0, & \text{otherwise} . \end{cases}$$

(b)

$$f = \begin{cases} x^2 - x, & y = 1,\ 0 < x < 1 . \\ 0, & \text{otherwise} . \end{cases}$$

Approximating functions: $\sinh \pi ny \sin \pi nx$, $n=1,2,\ldots$.

These functions are harmonic and satisfy the homogeneous boundary conditions. Because of the symmetry of the problem along the line $x = \frac{1}{2}$, the approximating functions used are $\sinh \pi(2n-1)y \sin \pi(2n-1)x$, $n=1,2,\ldots$.

Minimax error: See Table 6.6.

Table 6.6 Results for Square

	Minimax error ρ_n for function	
No. of functions	(a)	(b)
2	1.00	.0024
3	1.00	.0010
4	1.00	.00054
5	1.00	.00034
6	1.00	.00023
7	1.00	.00016
8	1.00	.00012
9	1.00	.00010

Exact solution for function (a) at (0.5,0.5): $u(0.5,0.5) = 0.25$ (see [22], p. 105).

Computed value of solution at (0.5,0.5) for n approximating functions: See Table 6.7.

Table 6.7 Computed solution for function (a) at (0.5,0.5)

n	value
2	.2474
3	.2519
4	.2505
5	.2504
6	.2504
7	.2504
8	.2455
9	.2501

6.1.4 Problem 4 For complete details, see Charmonman [2].

Boundary: Rectangle given by the lines $x = 0$, $x = -2$, $y = 0$, $y = -1$.

Boundary values:

$$f = \begin{cases} x, & y = -1, -1.5 < x < 0 . \\ 0, & \text{otherwise} . \end{cases}$$

Boundary-value problem: The mixed problem. The normal derivative of the solution is specified along $y = -1, -2 < x < -1.5$. Elsewhere, the value of the solution is given.

Approximating functions: i) $\sinh \frac{\pi n y}{-2} \sin \frac{\pi n x}{-2}$, $n=1,2,...$ where the value of the solution is given.

ii) $\frac{\pi n}{-2} \cosh \frac{\pi n y}{-2} \sin \frac{\pi n x}{-2}$, $n=1,2,...$ where the normal derivative of the solution is specified.

Minimax error: See Table 6.8.

Table 6.8 Results for Rectangle

No. functions	Minimax error ρ
2	.7461
3	.6317
4	.6140
5	.5875
6	.5264
7	.5204
8	.5052
9	.4612
10	.4598

6.2 Comments on Numerical Results

Problem 1: Davis [12] mentions this problem as a difficult test case because of the nonconvexity of the domain. Nevertheless, we were interested in testing our procedure for a fairly intricate region.

Boundary functions (a) to (e) yield a family of boundary values of increasing smoothness. The table of minimax errors show a decrease in the maximum boundary error as the boundary functions become smoother. In addition, the rate of decrease of the minimax error as we introduce more approximating functions is greater for smoother boundary functions.

Function (j) is harmonic over the region over which the solution is required and hence the boundary function is the solution to the Dirichlet problem. This test case was tried in [12] using a least-squares approach mentioned in Chapter I. We note that the maximum error along the boundary is smaller using our minimax procedure than the maximum error given by Davis' least-squares technique. However, the discrepancies in the interior seem to be larger for the minimax technique. In both approaches, the discrepancies in the interior are less than the computed maximum error along the boundary.

Problem 2: For test functions we selected four of the functions which were tried in Problem 1. (Boundary functions (a), (b), (c), (d) for the ellipse correspond to boundary functions (a), (b), (c), (h) for the bean.) For the family of functions (a) to (c) , we see that the minimax errors for the ellipse are less than the corresponding minimax errors for the bean. This is probably due to the nonconvexity of the bean-shaped region. On the other hand, the minimax errors for function (d) are bigger for the ellipse. This is a result of the singularity of the boundary function being closer to the ellipse than to the bean. For function (d) , we have listed the differences, for increasing numbers of approximating functions, between the true solution and the computed approximate solution at an interior point. The differences become smaller as the number of approximating functions increases. We note also that the discrepancies are well within the computed maximum error along the boundary.

Problem 3: The two boundary functions used show the difference in results between a case where the boundary function possesses a discontinuity on the boundary (function (a)) and a case where it does not (function (b)). The solution for the Dirichlet problem using function (a) can be written as an infinite series whose value is known at the interior point (0.5,0.5) . We have shown how the

computed value approaches the exact value at this point as more approximating functions are introduced. The approach to the true solution is quite uniform except for 8 approximating functions where the sudden divergence is something of a mystery, although it in no way contradicts any theory that we have presented.

Problem 4: Although we have not dealt with the mixed problem in the text of this thesis, we were interested to see how our minimax procedure would fare. For approximating the normal derivative of the solution along the boundary we have used a standard technique in numerical analysis: If L designates a linear operator and u a function, then we approximate $L(u)$ by L(approximation to u). It is interesting to note that the minimax error still decreases as the number of approximating functions increases. This leads us to believe that the technique may be used with success in other problem areas, such as the Neumann problem for Laplace's equation and problems involving multiply-connected domains. The large minimax errors are probably due to the singularity at the point $(-1.5,-1)$.

CHAPTER VII

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

7.1 Conclusions

The following conclusions have been drawn from the numerical computations:

1) As opposed to finite-difference techniques, the method presented is easy to apply to boundaries of irregular shape (for example, the bean-shaped region). Furthermore, obtaining the approximate solution at any desired point(s) in the interior of the region is relatively simple.

2) The technique is certainly as good as other collocation techniques such as the point-matching method and the least-squares method. The maximum error is usually a strictly decreasing function of the number of approximating harmonic functions. This is a virtue that the point-matching method does not always possess (see Section 1.3.1). The maximum error along the boundary is smaller for this method than for the least-squares method, and hence the bound of the error between the true solution and the computed solution is smaller. Furthermore, the maximum error is part of the automatic output of the routine, thus requiring no additional computation to obtain it.

3) The problems discussed show clearly the sensitivity of the method to both the boundary functions and the geometry of the boundary. The best results seem to come from boundary functions which are regular in a large portion of the plane. Poorer results are obtained for boundary data of low continuity and for boundary data which possess singularities near or on the boundary. The method prefers convex regions (such as the ellipse) over nonconvex regions (such as the bean), although convexity does not seem to play as important a role as the smoothness of the boundary functions.

7.2 Suggestions for Further Research

1) The heart of our computational algorithm is finding the minimax solution of an overdetermined, inconsistent system of linear equations. We have presented only one method of determining this minimax solution, namely the exchange algorithm. Cheney [3] briefly discusses two other methods - Polya's algorithm and the Descent algorithm. A paper* has come to our

* Bartels, R.H., and Golub, G.H., "Stable Numerical Methods for Obtaining the Chebyshev Solution to an Overdetermined System of Equations", JACM, 11, 1968, pp. 401-406.

attention as this thesis is in print which presents yet another technique. The method uses the exchange algorithm basically, but computes the inverse anew each iteration rather than updating the inverse from the previous iteration. The author claims the method to be more stable than the exchange algorithm as we have presented it. Furthermore, the algorithm requires only that the rank of the coefficient matrix be n, rather than that the rows of the coefficient matrix satisfy Haar's condition. A comparative analysis of these algorithms should be attempted.

2) Certainly, the Dirichlet problem for Laplace's equation, although important in its own right, is only one problem in the theory of partial differential equations. Numerical investigation of the mixed problem and the Neumann problem for Laplace's equation should be attempted. Other areas of investigation should include elliptic equations (of which Laplace's equation is a subset), boundary-value problems which involve multiply-connected domains, and the biharmonic equation.

3) We have given no attention to the use for which the solution of the Dirichlet problem is intended. Many problems require the integral of the solution over a region. For these problems, it might be better to minimize

$$M = \int_C \left| f(s) - \sum_{j=1}^{n} a_j g_j(s) \right| ds$$

rather than the maximum error along the boundary.

BIBLIOGRAPHY

1. Ahlfors, L.W., Complex Analysis, McGraw-Hill, Inc., New York, 1953.

2. Charmonman, S., "A Numerical Method of Solution of Free Surface Problems", Journal of Geophysical Research, 71, 1966, pp. 3861-3868.

3. Cheney, E.W., Introduction to Approximation Theory, McGraw-Hill, Inc., New York, 1966.

4. Cheng, K.C., "Analog Solution of Laminar Heat Transfer in Noncircular Ducts by Moiré Method and Point-Matching", J. Heat Transfer, Trans. ASME, 1966, pp. 175-181.

5. Conway, H.D., "The Approximating Analysis of Certain Boundary-Value Problems", J. Appl. Mech., 27, 1960, pp. 275-277.

6. Conway, H.D., "The Bending, Buckling, and Fleural Vibration of Simply Supported Polygonal Plates by Point-Matching", J. Appl. Mech., 28, 1961, pp. 288-291.

7. Conway, H.D., and Leissa, A.W., "A Method for Investigating Certain Eigenvalue Problems of the Buckling and Vibration of Plates", J. Appl. Mech., 27, 1960, pp. 557-558.

8. Conway, H.D., and Leissa, A.W., "Application of the Point-Matching Method to Shallow Spherical-Shell Theory", J. Appl. Mech., 29, 1962, pp. 745-747.

9. Curtiss, J.H., "Interpolation with Harmonic and Complex Polynomials to Boundary Values", J. Math. and Mech., 9, 1960, pp. 167-192.

10. Curtiss, J.H., "Interpolation by Harmonic Polynomials", J. Soc. Indust. Appl. Math., 10, 1962, pp. 709-736.

11. Curtiss, J.H., "Harmonic Interpolation in Fejér Points With the Faber Polynomials as a Basis", Math. Zeitschr., 86, 1964, pp. 75-92.

12. Davis, P.J., "Orthonormalizing Codes in Numerical Analysis", in Survey of Numerical Analysis, ed. J. Todd, McGraw-Hill, Inc., New York, 1962.

13. Davis, P.J., Interpolation and Approximation, Blaisdell, New York, 1963.

14. Davis, P.J., and Rabinowitz, P., "Advances in Orthonormalizing Computation", in Advances in Computers, 1961.

15. Forsythe, G., and Moler, C.B., Computer Solution of Linear Algebraic Systems, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1967.

16. Forsythe, G., and Wasow, W.R., Finite-Difference Methods for Partial Differential Equations, John Wiley and Sons, Inc., New York, 1960.

17. Fox, L., Numerical Solution of Ordinary and Partial Differential Equations, Pergamon Press, Oxford, 1962.

18. Handscomb, D.C. (ed.), Methods of Numerical Approximation, Pergamon Press, Oxford, 1966.

19. Leissa, A.W., and Niedenfuhr, F.W., "A Study of the Cantilevered Square Plate Subjected to Uniform Loading", J. Aerospace Sci., 29, 1962, pp. 162-169.

20. Lo, C., "The Solution of Plane Harmonic and Biharmonic Boundary-Value Problems in the Theory of Elasticity", Ph.D. Thesis, 1964, Ohio State University.

21. Merriman, G.M., "On the Expansion of Harmonic Functions in Terms of Normal Orthogonal Harmonic Polynomials", Amer. Journ. of Math., 53, 1931, pp. 589-596.

22. Miller, H.S., Partial Differential Equations in Engineering Problems, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1953.

23. Nash, W.A., "Several Approximate Analyses of the Bending of a Rectangular Cantilever Plate by a Uniform Normal Pressure", J. Appl. Mech., 19, 1952, pp. 33-36.

24. O'Jalvo, I.W., and Linzer, F.D., "Improved Point-Matching Techniques", Quart. J. Mech. and Appl. Math., 18, 1965, pp. 41-56.

25. Pennisi, L.L., Elements of Complex Variables, Holt, Rinehart and Winston, New York, 1963.

26. Poritsky, H., and Danforth, C.E., "On the Torsion Problem", Proc. 3rd U.S. National Congress of Appl. Mech., 27, 1960, pp. 431-441.

27. Quon, D., and Leung, P.H., "The Use of Dual Linear Programming in Formulating Approximating Functions by Using the Chebyshev Criterion", A.I.Ch.E. Journ., 1966, pp. 596-598.

28. Sekiya, T., "An Approximate Solution in the Problem of Elastic Plates with an Arbitrary External Form and a Circular Hole", Proc. 5th Japan National Congress for Appl. Mech., 1955, pp. 95-98.

29. Shuleshko, P., "Comparative Analysis of Different Collocation Methods on the Basis of the Solution of a Torsion Problem", Australian J. Appl. Sci., 12, 1960, pp. 194-210.

30. Slater, J.C., "Electron Energy Bands in Metals", Physical Review, 45, 1934, pp. 794-801.

31. Sobczyk, A., "On the Curtiss Non-Singularity Condition in Harmonic Polynomial Interpolation", J. Soc. Indust. Appl. Math., 12, 1964, pp. 499-514.

32. Sparrow, E.M., and Haji-Sheikh, A., "Flow and Heat Transfer in Ducts of Arbitrary Shape with Arbitrary Thermal Boundary Conditions", J. Heat Transfer, Trans. ASME, 1965, pp. 1-7.

33. Thorne, C.J., "Square Plates Fixed at Points", J. Appl. Mech., 15, 1948, pp. 73-79.

34. Walsh, J.L., "The Approximation of Harmonic Functions by Harmonic Polynomials and by Harmonic Rational Functions", Bull. Amer. Math. Soc., 35, 1929, pp. 499-544.

35. Walsh, J.L., "On Interpolation to Harmonic Functions by Harmonic Polynomials", Proc. Nat. Acad. Sci. U.S.A., 18, 1932, pp. 514-517.

36. Walsh, J.L., "Solution of the Dirichlet Problem for the Ellipse by Interpolating Harmonic Polynomials", J. Math. and Mech., 9, 1960, pp. 193-196.

APPENDIX

LISTINGS OF FORTRAN IV SUBPROGRAMS

The following pages contain listings of the IBM System 360 Fortran IV subprograms used in our numerical experiments.

Subroutines DECOMP, SOLVE, IMPRUV, and INVERT have been adapted from a set of programs given in Forsythe and Moler [15]. Details can be found therein. Nevertheless, for completeness, we give a brief description of each program.

Subroutine DECOMP (N,A,UL,DET,ISGND) uses Gaussian elimination to find N-by-N triangular matrices L and U so that $LU = PA$, where PA is the matrix A with its rows interchanged. The matrix $L - I + U$, where I is the identity matrix, is stored in UL. DET is the determinant of A and ISGND is the sign of the determinant of A.

Subroutine SOLVE (N,UL,B,X) uses the LU factorization from DECOMP to find an approximate solution to a system of equations $AX = B$.

Subroutine IMPRUV (N,A,UL,B,X,DIGITS) requires a copy of the original matrix A, its LU decomposition, a right-hand side B, and the approximate solution X computed by SOLVE. It carries out the iterative improve-

ment process, until, if possible, X is nearly accurate to machine precision (about 6 digits on the IBM S/360/67).

Subroutine INVERT (N,A,AINV) uses DECOMP, SOLVE, and IMPRUV to calculate the inverse of A , which is then stored in AINV.

Subroutine MINMAX computes the minimax solution and the associated minimax error of an overdetermined, inconsistent system of linear equations $Ax = b$.

Subroutine REMES computes the vector $c = (c_1,\dots,c_n)$ that minimizes

$$M = \max_{(x,y)\varepsilon C} \left|f(x,y) - \sum_{j=1}^{n} a_j g_j(x,y)\right| .$$

Subroutine ERRMAX finds $(\bar{x},\bar{y})\varepsilon C$ such that

$$\left|f(\bar{x},\bar{y}) - \sum_{j=1}^{n} c_j g_j(\bar{x},\bar{y})\right| = \max_{(x,y)\varepsilon C} \left|f(x,y) - \sum_{j=1}^{n} c_j g_j(x,y)\right| .$$

Subprogram U computes $\sum_{j=1}^{n} c_j g_j(x,y)$.

For the sake of completeness, we have included the subprograms SGN(X), F(X,Y), G(K,X,Y), and BOUND(X) . SGN(X) is the program for the mathematical function sgn x .

F(X,Y) is the program for the function f specified along the boundary; G(K,X,Y) is the program for the approximating function g_k ; BOUND(X) is the program of the boundary under consideration. These last three programs are valid only for problem 2, function (d) .

Subroutine SING is used by other routines to indicate the occurrence of an error condition.

```
   SUBROUTINE DECOMP(N,A,UL,DET,ISGND)
   DIMENSION A(20,20),UL(20,20),SCALES(20),IPS(20)
   COMMON IPS
   DO 5 I=1,N
   IPS(I)=I
   ROWNRM=0.0
   DO 2 J=1,N
   UL(I,J)=A(I,J)
 2 ROWNRM=AMAX1(ROWNRM,ABS(UL(I,J)))
   IF(ROWNRM)3,4,3
 3 SCALES(I)=1.0/ROWNRM
   GO TO 5
 4 CALL SING(1)
   RETURN
 5 CONTINUE
   DET=1.0
   IS=1
   NM1=N-1
   DO 16 K=1,NM1
   BIG=0.0
   DO 11 I=K,N
   IP=IPS(I)
   SIZE=ABS(UL(IP,K)*SCALES(IP))
   IF(SIZE-BIG)11,11,10
10 BIG=SIZE
   IDXPIV=I
11 CONTINUE
   IF(BIG)13,12,13
12 CALL SING(2)
   RETURN
13 IF(IDXPIV-K)14,15,14
14 IS=-IS
   J=IPS(K)
   IPS(K)=IPS(IDXPIV)
   IPS(IDXPIV)=J
15 KP=IPS(K)
   PIVOT=UL(KP,K)
   DET=DET*PIVOT
   KP1=K+1
   DO 16 I=KP1,N
   IP=IPS(I)
   EM=-UL(IP,K)/PIVOT
   UL(IP,K)=-EM
   DO 16 J=KP1,N
16 UL(IP,J)=UL(IP,J)+EM*UL(KP,J)
   KP=IPS(N)
   DET=IS*DET*UL(KP,N)
   ISGND=SGN(DET)
   IF(UL(KP,N))19,18,19
18 CALL SING(2)
```

```
19 RETURN
   END
```

```
      SUBROUTINE SOLVE(N,UL,B,X)
      DIMENSION UL(20,20),B(20),X(20),IPS(20)
      COMMON IPS
      NP1=N+1
      IP=IPS(1)
      X(1)=B(IP)
      DO 2 I=2,N
      IP=IPS(I)
      IM1=I-1
      SUM=0.0
      DO 1 J=1,IM1
    1 SUM=SUM+UL(IP,J)*X(J)
    2 X(I)=B(IP)-SUM
      IP=IPS(N)
      X(N)=X(N)/UL(IP,N)
      DO 4 IBACK=2,N
      I=NP1-IBACK
      IP=IPS(I)
      IP1=I+1
      SUM=0.0
      DO 3 J=IP1,N
    3 SUM=SUM+UL(IP,J)*X(J)
    4 X(I)=(X(I)-SUM)/UL(IP,I)
      RETURN
      END
```

```
      SUBROUTINE IMPROV(N,A,UL,B,X,DIGITS)
      DIMENSION A(20,20),UL(20,20),B(20),X(20),R(20),DX(20)
      DOUBLE PRECISION SUM,XJ,AIJ
      EPS=1.0E-6
      ITMAX=12
      XNORM=0.0
      DO 1 I=1,N
    1 XNORM=AMAX1(XNORM,ABS(X(I)))
      IF(XNORM)3,2,3
    2 DIGITS=-ALOG10(EPS)
      GO TO 10
    3 DO 9 ITER=1,ITMAX
      DO 5 I=1,N
      SUM=0.0
      DO 4 J=1,N
      XJ=X(J)
      AIJ=A(I,J)
    4 SUM=SUM+AIJ*XJ
      SUM=B(I)-SUM
    5 R(I)=SUM
      CALL SOLVE(N,UL,R,DX)
      DXNORM=0.0
      DO 6 I=1,N
      T=X(I)
      X(I)=X(I)+DX(I)
      DXNORM=AMAX1(DXNORM,ABS(X(I)-T))
    6 CONTINUE
      IF(ITER-1)8,7,8
    7 DIGITS=-ALOG10(AMAX1(DXNORM/XNORM,EPS))
    8 IF(DXNORM-EPS*XNORM)10,10,9
    9 CONTINUE
      CALL SING(3)
   10 RETURN
      END
```

```
      SUBROUTINE INVERT(N,A,AINV)
      DIMENSION A(20,20),AINV(20,20),B(20),UL(20,20),X(20)
      CALL DECOMP(N,A,UL,DET,ISGND)
      DO 3 J=1,N
      DO 1 I=1,N
      B(I)=0.0
      IF(I.EQ.J)B(I)=1.0
    1 CONTINUE
      CALL SOLVE(N,UL,B,X)
      CALL IMPROV(N,A,UL,B,X,DIGITS)
      DO 2 I=1,N
    2 AINV(I,J)=X(I)
    3 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE MINMAX(M,N,A,B,IV,X)
C
C  M IS THE NUMBER OF EQUATIONS.
C  N IS THE NUMBER OF UNKNOWNS.
C  A IS THE M-BY-N MATRIX OF COEFFICIENTS.
C  B IS THE RIGHT-HAND SIDE.
C  M,N,A,B MUST BE SUPPLIED BY THE CALLING PROGRAM.
C  IV IS THE VECTOR OF INDICES OF THE EQUATIONS
C      PRESENTLY UNDER CONSIDERATION. AN INITIAL
C      GUESS MUST BE SUPPLIED BY THE CALLING
C      PROGRAM. ON RETURN,IV CONTAINS THE INDICES
C      OF THE EQUATIONS OF THE CRITICAL SUBSYSTEM.
C  X IS AN (N+1)-COMPONENT VECTOR.ON RETURN,
C      X(1),...X(N) CONTAIN THE MINIMAX SOLUTION,
C      AND X(N+1) CONTAINS THE MINIMAX ERROR.
C
      DIMENSION A(100,19),B(100),IV(20),X(20)
      DIMENSION AA(20,20),D(20,20),UL(20,20)
      REAL S(20),RHO(100),LAMDA(20)
      DOUBLE PRECISION LKJ,DJK,AIJ,BIR,XJ,LJ,SUM
C
C              STEP 1
C
      NP1=N+1
      DO 10 K=1,NP1
      DO 5 I=1,N
      IF(I-K)1,2,2
    1 IR=IV(I)
      GO TO 3
    2 IR=IV(I+1)
    3 DO 4 J=1,N
    4 AA(I,J)=A(IR,J)
    5 CONTINUE
      CALL DECOMP(N,AA,UL,DET,ISGND)
C
C              STEP 2
C
   10 S(K)=ISGND*(-1)**K
C
C              STEP 3
C
      DO 20 I=1,NP1
      IR=IV(I)
      DO 15 J=1,N
   15 AA(I,J)=A(IR,J)
   20 AA(I,NP1)=S(I)
      CALL INVERT(NP1,AA,D)
C
C              STEP 4
C
```

```
   25 DO 30 K=1,NP1
      SUM=0.0
      DO 29 J=1,NP1
      IR=IV(J)
      DKJ=D(K,J)
      BIR=B(IR)
   29 SUM=SUM+DKJ*BIR
   30 X(K)=SUM
C
C              STEP 5
C
      IALPHA=1
      RHOMAX=0.0
      DO 40 I=1,M
      SUM=0.0
      DO 35 J=1,N
      AIJ=A(I,J)
      XJ=X(J)
   35 SUM=SUM+AIJ*XJ
      RHO(I)=B(I)-SUM
C
C              STEP 6
C
      IF(ABS(RHO(I))-RHOMAX)40,40,39
   39 RHOMAX=ABS(RHO(I))
      IALPHA=I
   40 CONTINUE
C
C              STEP 7
C
      IF(RHOMAX-X(NP1))71,71,42
   42 DO 43 K=1,NP1
      IF(IALPHA-IV(K))43,71,43
   43 CONTINUE
C
C              STEP 8
C
      U=SGN(RHO(IALPHA))
C
C              STEP 9
C
      RMAX=-1E75
      IK=1
      DO 50 K=1,NP1
      SUM=0.0
      DO 45 J=1,N
      DJK=D(J,K)
   45 SUM=SUM+A(IALPHA,J)*DJK
      LAMDA(K)=SUM+U*D(NP1,K)
      IF(D(NP1,K))47,55,47
```

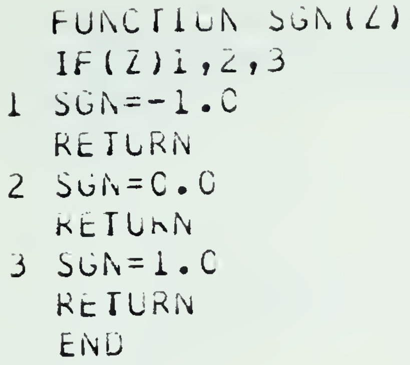
```
C
C               STEP 10
C
   47 T=LAMDA(K)*U/D(NP1,K)
      IF(T-RMAX)50,50,49
   49 RMAX=T
      IK=K
   50 CONTINUE
      IF(LAMDA(IK))59,55,59
   55 CALL SING(4)
C
C               STEP 11
C
   59 DO 60 K=1,NP1
   60 D(K,IK)=D(K,IK)/DBLE(LAMDA(IK))
      DO 70 K=1,NP1
      DO 70 J=1,NP1
      IF(J-IK)69,70,69
   69 DKJ=D(K,J)
      LJ=LAMDA(J)
      D(K,J)=DKJ-LJ*D(K,IK)
   70 CONTINUE
C
C               STEP 12
C
      IV(IK)=IALPHA
      GO TO 25
   71 RETURN
      END
```

```
      SUBROUTINE REMES(MM,N,PTS,IV,C)
C
C   MM IS THE NUMBER OF POINTS IN THE INITIAL DISCRETE
C        POINT SET.
C   N IS THE NUMBER OF APPROXIMATING FUNCTIONS.
C   PTS IS THE VECTOR OF ABSCISSAS OF THE POINTS IN
C        THE DISCRETE POINT SET.
C   IV IS THE VECTOR OF INDICES OF POINTS IN THE
C        CRITICAL SUBSYSTEM RETURNED BY MINMAX.
C   C IS AN (N+1)-COMPONENT VECTOR.C(1),...C(N)
C        CONTAIN THE MINIMIZING COEFFICIENTS. C(N+1)
C        CONTAINS THE ASSOCIATED MINIMAX ERROR.
C
      DIMENSION PTS(100),IV(20),C(20),A(100,19),B(100)
      M=MM
      EPS=1.0E-3
      ITMAX=90
C
C   EPS AND ITMAX ARE CHOSEN BY THE PROGRAMMER.
C   SET UP THE MATRIX A AND THE VECTOR B FOR MINMAX.
C
      NP1=N+1
      DO 5 I=1,M
      X=PTS(I)
      Y=BOUND(X)
      B(I)=F(X,Y)
      DO 5 J=1,N
    5 A(I,J)=G(J,X,Y)
      DO 25 ITER=1,ITMAX
C
C   SELECT COEFFICIENTS TO MINIMIZE MAXIMUM ERROR
C        ON DISCRETE POINT SET PRESENTLY BEING
C        CONSIDERED.
C
      CALL MINMAX(M,N,A,B,IV,C)
C
C   FIND POINT (CALLED XMAX) OF MAXIMUM ERROR
C        (CALLED EMAX) USING COEFFICIENTS CALCULATED
C        FROM PREVIOUS STATEMENT.
C
      CALL ERRMAX(N,C,XMAX,EMAX)
C
C   EXIT ROUTINE.
C
      IF(EMAX-C(NP1))50,50,15
   15 IF((EMAX-C(NP1))/EMAX-EPS)50,50,16
C
C   INTRODUCE POINT OF MAXIMUM ERROR INTO DISCRETE
C        POINT SET. UPDATE A AND B.
C
```

```
16 M=M+1
   PTS(M)=XMAX
   YMAX=BOUND(XMAX)
   B(M)=F(XMAX,YMAX)
   DO 20 J=1,N
20 A(M,J)=G(J,XMAX,YMAX)
25 CONTINUE
   CALL SING(5)
50 RETURN
   END
```

```
      SUBROUTINE ERRMAX(N,C,XMAX,EMAX)
C
C   A AND B DENOTE THE END POINTS OF THE RANGE OF THE
C        INDEPENDENT VARIABLE. THEY MUST BE SUPPLIED
C        BY THE PROGRAMMER.
C
      DIMENSION C(20)
      PI=3.141593
      N20=20*N
      EMAX=0.0
      DO 10 K=1,N20
C
C   THE KTH ZERO OF THE CHEBYSHEV POLYNOMIAL OF
C        DEGREE 20*N,NORMALIZED TO THE INTERVAL (A,B).
C
      X=0.5*((B-A)*COS(PI*(K-0.5)/N20)+B+A)
      Y=BCOND(X)
      E=ABS(F(X,Y)-U(N,C,X,Y))
      IF(EMAX-E)5,10,10
    5 EMAX=E
      XMAX=X
   10 CONTINUE
      RETURN
      END
```

```
      FUNCTION U(N,CFS,X,Y)
      DIMENSION CFS(20)
      DOUBLE PRECISION SUM,CK
      SUM=0.0
      DO 10 K=1,N
      CK=CFS(K)
   10 SUM=SUM+CK*G(K,X,Y)
      U=SUM
      RETURN
      END
```



```
      FUNCTION SGN(Z)
      IF(Z)1,2,3
    1 SGN=-1.0
      RETURN
    2 SGN=0.0
      RETURN
    3 SGN=1.0
      RETURN
      END
```

```
FUNCTION F(X,Y)
F=ALOG((X+1.5)**2+Y**2)
RETURN
END
```

```
      FUNCTION G(K,X,Y)
      COMPLEX Z
      KM1=K-1
      IF(KM1)1,1,2
    1 G=1.0
      RETURN
    2 Z=CMPLX(X,Y)
      G=REAL(Z**KM1)
      RETURN
      END
```

```
FUNCTION BOUND(X)
BOUND=0.5*SQRT(1-X**2)
RETURN
END
```

```
      SUBROUTINE SING(IWHY)
      GO TO (1,2,3,4,5),IWHY
    1 WRITE(6,11)
      RETURN
    2 WRITE(6,12)
      RETURN
    3 WRITE(6,13)
      RETURN
    4 WRITE(6,14)
      RETURN
    5 WRITE(6,15)
      RETURN
   11 FORMAT('0 MATRIX WITH ZERO ROW IN DECOMP.')
   12 FORMAT('0 SINGULAR MATRIX IN DECOMP.')
   13 FORMAT('0 NO CONVERGENCE IN IMPROV.')
   14 FORMAT('0 HAAR CONDITION VIOLATED IN MINMAX.')
   15 FORMAT('0 NO CONVERGENCE IN REMES.')
      END
```

Minimax Coefficients for the Bean Problem

$$f(x,y) = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases}$$

n = 3 :	.132231	.461538	.092308		
n = 5 :	.093404	.486649	.008827	.608221	-.034363
n = 7 :	.092597	.536986	-.105155	.653991	.018744
	.180471	.441858			
n = 19:	.075025	.514878	-.209529	.989064	.015350
	-.224194	1.03329	-2.27918	-.035892	.213942
	-2.60345	6.39725	-.026738	1.27328	4.87422
	-9.69454	-.596876	-11.2224	-5.88364	

$$f(x,y) = \begin{cases} 0\ , & x \le 0 \\ x^3 & x > 0 \end{cases}$$

n = 3 :	.031894	.120151	-.000872		
n = 5 :	.017586	.014766	-.014176	.199221	-.006899
n = 7 :	.013713	.089248	-.032558	.196624	.034272
	.212272	.065763			
n = 19:	.010926	.079131	-.048096	.236794	-.024109
	.241918	.200170	-.175985	.255733	-.415811
	-.305695	.488001	-.453883	1.14545	.212120
	-1.02356	.561388	-1.88455	.951538	

$$f(x,y) = \begin{cases} 0\ , & x \le 0 \\ x^5, & x > 0 \end{cases}$$

```
n = 3 :    .007117   .032789    0.0
n = 5 :    .003361   .023898   -.002281   .056748   .005689
n = 7 :    .002812   .021016   -.007908   .049424   .011138
           .055872   .019046
n = 19:    .002201   .016757   -.009721   .055003  -.004772
           .072625   .045666    .002280   .071505  -.066786
          -.035986   .078506   -.096603   .184970   .009972
          -.229025   .153825   -.347786   .244229
```

$$f(x,y) = e^x \cos y + \log[(1-y)^2 + x^2]$$

```
n = 3 :   1.19201    .861679   -1.62099
n = 5 :   1.04665    .993063   -1.80887   1.40812   -.222934
n = 7 :   1.01423    .991961   -1.94136   1.36212   -.017493
           .309424   .610016
n = 19:   1.00007    .999835   -1.99968   1.49914   -.000545
           .168320   .664599   -.453063   .004362  -.000485
          -.388996   .309699   -.018050   .039101   .243903
          -.168254   .007315   -.114652  -.120182
```